IEEE HOME | SEARCH IEEE | SHOP | WEB ACCOUNT | CONTACT IEEE

IEEE

Membership Publications/Services Standards Conferences Careers/Jobs

IEEE Xplore® RELEASE 1.8

Welcome
United States Patent and Trademark Office

Help FAQ Terms IEEE Peer Review

Quick Links

» ABS

Welcome to IEEE Xplore®

○ Home
○ What Can I Access?
○ Log-out

Tables of Contents

○ Journals & Magazines
○ Conference Proceedings
○ Standards

Search

○ By Author
○ Basic
○ Advanced

Member Services

○ Join IEEE
○ Establish IEEE Web Account
○ Access the IEEE Member Digital Library

IEEE Enterprise

○ Access the IEEE Enterprise File Cabinet

Print Format

Search Results [PDF FULL-TEXT 648 KB] PREV NEXT DOWNLOAD CITATION

Request Permissions RIGHTSLINK

Learning to identify and track faces in image seque

Edwards, G.J. Taylor, C.J. Cootes, T.F.
Dept. of Med. Biophys., Manchester Univ., UK;

This paper appears in: **Computer Vision, 1998. Sixth International Confe**

Meeting Date: 01/04/1998 - 01/07/1998
Publication Date: 4-7 Jan. 1998
Location: Bombay India
On page(s): 317 - 322
Reference Cited: 9
Number of Pages: 1164
Inspec Accession Number: 6015952

Abstract:
We address the problem of robust face identification in the presence of pose, and expression variation. Previous approaches to the problem have assumed models of variation for each individual, estimated from pooled training data. W a method of updating a first order global estimate of identity by learning the c specific correlation between the estimate and the residual variation during a s This is integrated with an optimal tracking scheme, in which identity variation decoupled from pose, lighting and expression variation. The method results in tracking and a more stable estimate of facial identity under changing condition

Index Terms:
face recognition image sequences expression variation first order global estimate im sequences lighting optimal tracking scheme pose robust face identification

Documents that cite this document
Select link to view other documents in the database that cite this one.

Search Results [PDF FULL-TEXT 648 KB] PREV NEXT DOWNLOAD CITATION

Home | Log-out | Journals | Conference Proceedings | Standards | Search by Author | Basic Search | Advanced Search | Join IEEE | Web Account | New this week | OPAC Linking Information | Your Feedback | Technical Support | Email Alerting | No Robots Please | Release Notes | IEEE Online Publications | Help | FAQ| Terms | Back to Top

Learning to Identify and Track Faces in Image Sequences

G.J. Edwards, C.J. Taylor and T.F. Cootes
Dept. Medical Biophysics,
Manchester University, UK
email: gje@sv1.smb.man.ac.uk

Abstract

We address the problem of robust face identification in the presence of pose, lighting, and expression variation. Previous approaches to the problem have assumed similar models of variation for each individual, estimated from pooled training data. We describe a method of updating a first order global estimate of identity by learning the class-specific correlation between the estimate and the residual variation during a sequence. This is integrated with an optimal tracking scheme, in which identity variation is decoupled from pose, lighting and expression variation. The method results in robust tracking and a more stable estimate of facial identity under changing conditions.

1 Introduction

Locating and interpreting faces in images and image sequences is a difficult problem in machine vision, due to the inherent variability between and within individuals. The appearance of a face in an image varies with the identity of the individual, pose, lighting conditions, and deformations due to expression or speech. Previous work has shown how the problem can be addressed by using statistical models which combine shape and intensity variation within a single framework. These *Combined Appearance Models* [4], account for all sources of variability in face images. We are interested in isolating the specific sources of variation present in face images, in order to improve identity recognition in the presence of pose, lighting and expression variation, and to allow more robust tracking, by modelling the dynamics of different sources of variability separately. We show how a discriminant analysis method [4] can be used to acheive this to a first-order approximation by assuming the sources of variation are orthogonal and identical for different individuals. This last assumption is necessary because it is unrealistically restrictive to assume a sufficiently large training set for every individual, to determine a class-specific model of variability. We describe how, using image sequences, the first-order approximation to the separation of sources of variability can be improved with a class-specific correction, to give a class-specific representation for particular individuals. This allows a more precise description of identity, and better decoupling of the sources of variation. The decoupling is used to provide separate dynamic models of variation for sequences which can be used in a Kalman filtering framework. We show an example of the method used to track a face in an image sequence, acheiving robust tracking, and yielding a more precise estimate of identity.

2 Background

In many face recognition applications the task is to locate faces in images, and identify them in a way which is robust with respect to changes in pose, expression, and lighting conditions. In this section we outline briefly an existing model-based approach to location and recognition, on which the current work is based.

2.1 Statistical Models

Statistical modelling of facial appearance has proved a successful approach to coding and interpreting face images, also providing a useful basis for locating faces in images. Kirby and Sirovich [7] describe a compact representation of facial appearance, where face images are decomposed into weighted sums of basis images using a Karhumen-Loeve expansion. The patch containing the face is coded using 50 expansion coefficients from which an approximation to the original can be reconstructed. Turk and Pentland [9] describe face identification using this 'Eigenface' representation. Lanitis et al. [8] describe the representation of both face shape and grey-level appearance; they use a *Point Distribution Model* (PDM) [3] to describe shape and an approach similar to Kirby and Sirovich [7] to represent shape-normalised grey-level appearance. More recently, Edwards et al. [4] have described the combination of shape and grey-level variation within a single statistical appearance model, which they call a *Combined Appearance Model.*

Welcome to IEEE Xplore®

- ○ Home
- ○ What Can I Access?
- ○ Log-out

Tables of Contents

- ○ Journals & Magazines
- ○ Conference Proceedings
- ○ Standards

Search

- ○ By Author
- ○ Basic
- ○ Advanced

Member Services

- ○ Join IEEE
- ○ Establish IEEE Web Account
- ○ Access the IEEE Member Digital Library

IEEE Enterprise

- ○ Access the IEEE Enterprise File Cabinet

Print Format

Search Results [PDF FULL-TEXT 1444 KB] PREV DOWNLOAD CITATION

Request Permissions RIGHTSLINK

Illumination cones for recognition under variable li faces

Georghiades, A.S. Kriegman, D.J. Belhurneur, P.N.
Center for Comput. Vision & Control, Yale Univ., New Haven, CT, USA;

This paper appears in: **Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on**

Meeting Date: 06/23/1998 - 06/25/1998
Publication Date: 23-25 June 1998
Location: Santa Barbara, CA USA
On page(s): 52 - 58
Reference Cited: 17
Number of Pages: xvii+970
Inspec Accession Number: 5985824

Abstract:

Due to illumination variability, the same object can appear dramatically differe when viewed in fixed pose. To handle this variability, an object recognition sy employ a representation that is either invariant to, or models this variability. T presents an appearance-based method for modeling the variability due to illum the images of objects. The method differs from past appearance-based metho however, in that a small set of training images is used to generate a represen illumination cone-which models the complete set of images of an object with L reflectance map under an arbitrary combination of point light sources at infini method is both an implementation and extension (an extension in that it mod shadows) of the illumination cone representation proposed in Belhumeur and (1996). The method is tested on a database of 660 images of 10 faces, and th exceed those of popular existing methods

Index Terms:

face recognition learning (artificial intelligence) lighting object recognition Lambertia map faces illumination variability object recognition training images

Documents that cite this document

Select link to view other documents in the database that cite this one.

Illumination Cones for Recognition Under Variable Lighting: Faces *

Athinodoros S. Georghiades David J. Kriegman Peter N. Belhumeur

Center for Computational Vision and Control
Yale University
New Haven, CT 065204267

Abstract

Due to illumination *variability, the* same *object* can appear *dramatically different even* when *viewed* in *fixed pose. To handle this variability, an object recognition system must employ* a *representation that is either invariant to, or models this variability. This paper* presents an *appearance-based method for modeling the variability due to illumination in the images of objects. The method differs from past appearance-based methods, however, in that* a *small set of training images is used to* generate *a representation the illumination cone which models the* complete set of images *of an object with Lambertian* reflectance *map under* an *ar*bitrary *combination of point light sources at infinity. This method is both* an *implementation and extension* (an *extension in that it models cast shadows) of the illumination cone representation proposed in [3]. The method is tested on* a database *of 660* images *of 10 faces,* and *the* results *exceed those of popular existing methods.*

1 Introduction

An object's appearance depends in large part on the way in which it is viewed. Often slight changes in pose and illumination produce large changes in an object's appearance. While there has been a great deal of literature in computer vision detailing methods for handling image variation produced by changes in pose, few efforts have been devoted to image variation produced by changes in illumination. For the most part, object recognition algorithms have either ignored illumination variation, or dealt with it by measuring some property or feature of the image – e.g., edges or corners – which is, if not invariant, at least insensitive to this variability. Yet, edges and corners do not contain all of the information useful for recognition. Furthermore, objects which are not simple polyhedra or are not composed of piecewise constant albedo patterns often produce inconsistent edge and corner maps.

Methods have recently been introduced which use low-dimensional representations of images of objects to perform recognition, see for example [5, 11, 16]. These methods, often termed appearance-based methods, differ from the feature-based methods mentioned above in that their low-dimensional representation is, in a least-squared sense, faithful to the original image. Systems such as SLAM [11] and Eigenfaces [16] have demonstrated the power of appearance-based methods both in ease of implementation and in accuracy. Yet these methods suffer from an important drawback: recognition of an object (or face) under a particular pose and lighting can be performed reliably *provided that object has* been previously seen *under similar circumstances*. In other words, these methods in their original form have no way of extrapolating to novel viewing conditions.

The "illumination cone" method of [3] is, in spirit, an appearance-based method for recognizing objects under extreme variability in illumination. However, the method differs substantially from previous methods in that a small number of images of each object under small changes in lighting is used to generate a representation, the illumination cone, of all images of the object (in fixed pose) under all variation in illumination. This paper focuses on issues for building the illumination cone representation from training images and using it for recognition.

While the structure of the set of images under variable illumination was characterized in [3] and the relevant results are summarized in Sec. 2, no methods for performing recognition were presented. In this paper, such recognition algorithms are introduced. Furthermore, the cone representation is extended to explicitly model cast shadows produced by objects which have non-convex shapes. This extension is non-trivial, requiring that the surface normals for the objects be recovered up to a *shadow preserving* generalized bas-relief (GBR) transformation.

The effectiveness of these algorithms and the cone representation are validated within the context of face recognition – it has been observed by Moses, Adini and Ullman that the variability in an image due to illumination is often greater than that due to a change in the person's identity [10]. Figure 1 shows the variability for a single individual. It has been observed that methods for face recognition based on finding local image features and using their geometric relation are generally ineffective [4]. Hence, faces provide an interesting and useful class of objects for testing the power of the illumination cone representation.

In this paper, we empirically compare these new methods to a number of popular techniques such as correlation [4] and Eigenfaces [9,16] as well as more

*A. S. Georghiades and D. J. Kriegman were supported by NSF under NYI IRI-9257990. P. N. Relhumeur was supported by a Presidential Early Career Award, an NSF Career Award IRI-9703134, and ARO grant DAAH04-95-1-0494.

recently developed techniques such as distance to linear subspace [2, 5, 12, 13]; the latter technique has been shown to be much less sensitive to illumination variation than the former. However, these methods also break down as shadowing becomes very significant. As we will see, the presented algorithm based on the illumination cone outperforms all of these methods on a database of 660 images. It should be noted that our objective in this work is to focus solely on the issue of illumination variation whereas other approaches have been more concerned with issues related to large image databases, face finding, pose, and facial expressions.

2 The Illumination Cone

In earlier work, it was shown that for an object with convex shape and Lambertian reflectance, the set of all images under an arbitrary combination of point light sources forms a convex polyhedral cone in the image space $\mathbb{R}^n$. This cone can be constructed from as few as three images [3]. Here we summarize the relevant results.

To begin, consider a convex object with a Lambertian reflectance function which is illuminated by a single point source at infinity. Let $\mathbf{x} \in \mathbb{R}^n$ denote an image of this object with n pixels. Let $B \in \mathbb{R}^{n \times 3}$ be a matrix where each row of B is the product of the albedo with the inward pointing unit normal for a point on the surface projecting to a particular pixel in the image. A point light source at infinity can be represented by $\mathbf{s} \in \mathbb{R}^3$ signifying the product of the light source intensity with a unit vector in the direction of the light source. A convex Lambertian surface with normals and albedo given by B, illuminated by $\mathbf{s}$, produces an image $\mathbf{x}$ given by

$$\mathbf{x} = \max(B\mathbf{s}, \mathbf{0}), \qquad (1)$$

where $\max(B\mathbf{s}, \mathbf{0})$ sets to zero all negative components of the vector $B\mathbf{s}$. The pixels set to zero correspond to the surface points lying in an *attached shadow*. Convexity of the object's shape is assumed at this point to avoid *cast shadows* (shadows that the object casts on itself). While attached shadows are defined by local geometric condition, cast shadows must satisfy a global condition.

When no part of the surface is shadowed, $\mathbf{x}$ lies in the 3-D subspace $\mathcal{L}$ given by the span of the matrix B. It can be shown that the subset $\mathcal{L}_0 \subset \mathcal{L}$ having no shadows (i.e., falling in the non-negative orthant[1]) forms a convex cone [3].

The illumination subspace $\mathcal{L}$ slices through other orthants as well as the non-negative orthant. Let $\mathcal{L}_i$ be the intersection of the illumination subspace $\mathcal{L}$ with an orthant i in $\mathbb{R}^n$ through which $\mathcal{L}$ passes. Certain components of $\mathbf{x} \in \mathcal{L}_i$ are always negative and others always greater than or equal to zero. Since image intensity is always non-negative, the image corresponding to points in $\mathcal{L}_i$ is formed by the projection P_i given by Equation 1. The projection P_i is such that it leaves the non-negative components of $\mathbf{x} \in \mathcal{L}_i$ untouched, while the negative components of $\mathbf{x}$ become zero. The projected set $P_i(\mathcal{L}_i)$ is also a convex cone. $\mathcal{L}$ intersects at most $n(n-1)+2$ orthants [3], and so the set of images created by varying the direction and strength of a *single* light source at infinity is given by the union of at most $n(n-1)+2$ convex cones, each of which is at most three dimensional.

If an object is illuminated by k light sources at infinity, then the image is given by the superposition of the images which would have been produced by the individual light sources, i.e.,

$$\mathbf{x} = \sum_{i=1}^{k} \max(B\mathbf{s}_i, \mathbf{0}) \qquad (2)$$

where $\mathbf{s}_i$ is a single light source. It follows that the set of all possible images $\mathcal{C}$ of a convex Lambertian surface created by varying the direction and strength of an arbitrary number of point light sources at infinity is a convex cone.

Furthermore, it is shown in [3] that any image in the cone $\mathcal{C}$ (including the boundary) can be found as a convex combination of *extreme rays* given by

$$\mathbf{x}_{ij} = \max(B\mathbf{s}_{ij}, \mathbf{0}), \qquad (3)$$

where

$$\mathbf{s}_{ij} = \mathbf{b}_i \times \mathbf{b}_j. \qquad (4)$$

The vectors $\mathbf{b}_i$ and $\mathbf{b}_j$ are the rows of B with $i \neq j$. It is clear that there are at most $m(m-1)$ extreme rays (images) for $m \leq n$ independent surface normals. Since there are a finite number of extreme rays, the convex cone is polyhedral.

3 Constructing the Illumination Cone

Equations 3 and 4 suggest a way to construct the illumination cone for each individual: gather three or more images of the face under varying illumination without shadowing and use these images to estimate the three-dimensional illumination subspace $\mathcal{L}$. One way of estimating this is to normalize the images to be of unit length, and then use singular value decomposition (SVD) to estimate the best three-dimensional orthogonal basis B^* in a least square sense. Note that the basis B^* differs from B by an unknown linear transformation, i.e., $B = B^*A$ where $A \in GL(3)$; for any light source, $\mathbf{x} = B\mathbf{s} = (B^*A)(A^{-1}\mathbf{s})$. Nonetheless from B^*, the extreme rays defining the illumination cone $\mathcal{C}$ can be computed using Equations 3 and 4. This method, introduced in [3], was named the *illumination subspace method*.

[1] By orthant we mean the high-dimensional analogue to quadrant, i.e., the set $\{\mathbf{x} | \mathbf{x} \in \mathbb{R}^n$, with certain components of $\mathbf{x} \geq 0$ and the remaining components of $\mathbf{x} < 0\}$. By non-negative orthant we mean the set $\{\mathbf{x} | \mathbf{x} \in \mathbb{R}^n$, with all components of $\mathbf{x} \geq 0\}$.

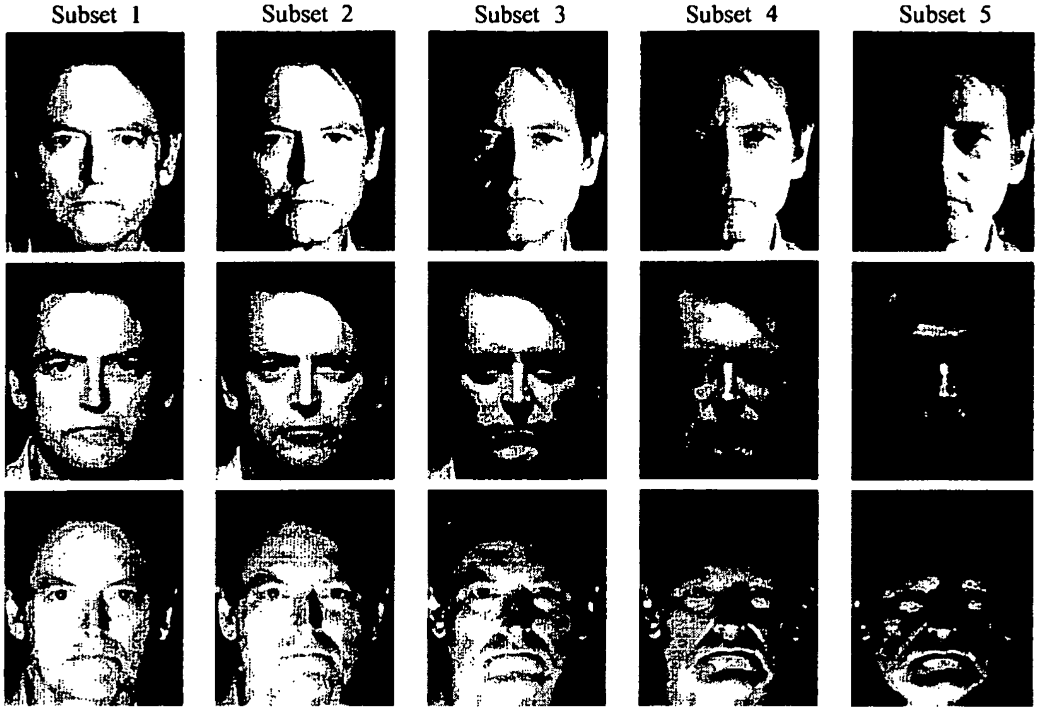


Figure 1: Example images from each subset of the Harvard Database used to test the algorithms

The first problem that arises with the above procedure is with the estimation of $\boldsymbol{B}'$. For even a convex object whose Gaussian image covers the Gauss sphere, there is only one light source direction (the viewing direction) for which no point on the surface is in shadow. For any other light source direction, shadows will be present. For faces, which are not convex, shadowing in the modeling images is likely to be more pronounced. When SVD is used to estimate B^* from images with shadows, these systematic errors can bias the estimate of B^* significantly. Therefore, alternative ways are needed to estimate B^* that take into account the fact that some data values should not be used in the estimation.

The next problem is that usually $\boldsymbol{m}$, the number of independent normals in $\boldsymbol{B}$, can be large (more than a thousand) hence the number of extreme rays needed to completely define the illumination cone can run in the millions. Therefore, we must approximate the cone in some fashion; in this work, we choose to use a small number of extreme rays (images). In [3] it was shown empirically that the cone is flat (i.e., elements lie near a low dimensional linear subspace), and so the hope is that a sub-sampled cone will provide an approximation that leads to good recognition performance. In our experience, around 60-80 images are sufficient, provided that the corresponding light source directions $\mathbf{s}_{ij}$ are more or less uniform on the illumination sphere. The resulting cone C* is a subset of the object's true cone C. An alternative approximation to C can be obtained by directly sampling the space of light source directions rather than generating the samples through Eq. 4. While the resulting images form the extreme rays of the representation C^* and lie on the boundary of C, they are not necessarily extreme rays of C. Again C^* is a subset of C.

The last problem comes from the fact that faces arc non-convex, and so cast shadows cover significant portions of the face under extreme illumination (See the images from Subsets 4 and 5 in Fig. 1). The image formation model (Eq. 1) used to develop the illumination cone does not account for cast shadows. For the light source directions of the extreme rays given by Equation 4, we need to predict which pixels will be in cast shadows.

It has been shown [1,17] that from multiple images where the light source directions are unknown, one can only recover a Lambertian surface up to a three-parameter family given by the generalized bas-relief (GBR) transformation. This family scales the relief (flattens or extrudes) and introduces an additive plane. Consequently, when computing $\mathbf{s}^*_{ij}$ from B^*, the light source direction differs from the true light source by a GBR transformation. Since shadows are preserved under these transformation [1], images synthesized from a surface whose normal field is given by B^* under light source $\mathbf{s}^*_{ij}$ will have correct shadowing. Thus, in constructing the extreme rays of the cone, we first reconstruct a surface (a height function) and then use ray-tracing techniques to determine which points lie in a cast shadow. It should be noted that the vector field B^* estimated via SVD may not be integrable, so prior to reconstructing the surface up to GBR, integrability of $\boldsymbol{B}^*$ is enforced.

This leads to the following steps for constructing a representation of the illumination cone of an individual from a set of images taken under unknown lighting. Details of these steps are given below.

1. Estimate B^* from training images.
2. Enforce integrability of B^*.
3. Reconstruct the surface up to GBR.
4. For a set of light source directions that uniformly sample the sphere, synthesize extreme rays (images) of the cone that account for cast and attached shadows.

3.1 Estimating B^*

Using singular value decomposition directly on the images leads to a biased estimate of $\boldsymbol{B}^*$ due to shadows. In addition, portions of some of the images from the Harvard database used in our experiments were saturated. Both shadows formed under a single light source and saturations can be detected by thresholding and labeled as "missing" - these pixels do not satisfy the linear equation x = $\boldsymbol{Bs}$. Thus, we need to estimate the 3-D linear subspace $\boldsymbol{B}^*$ from images with missing values.

Define the data matrix for c images of an individual to be X = $[\mathbf{x}_1 \ldots \mathbf{x}_c]$. If there were no shadowing, X would be rank 3, and we could use SVD to decompose X into X = B^*S^* where S* is a 3 x c matrix of the light source direction for all c images. To estimate a basis B^* for the 3-D linear subspace $\mathcal{L}$ from image data with missing elements, we have implemented a variation of [14]; see also [15,8].

The overview of this method is as follows: without doing any row or column permutations sift out all the full rows (with no invalid data) of matrix X to form a full sub-matrix $\tilde{X}$. Perform SVD on $\tilde{X}$ and get an initial estimate of S^*. Fix S* and estimate each of the rows of B^* independently using least squares. Then, fix B^* and estimate each of the light source direction $\mathbf{s}_i$ independently. Repeat last two steps until estimates converge. The inner workings of the algorithm are given as follows: Let $\mathbf{b}_i$ be the ith row of B^*, let $\mathbf{x}_i$ be the ith row of X. Let p be the indices of non-missing elements in $\mathbf{x}_i$, and let $\mathbf{x}_i^p$ be the row obtained by taking only the non-missing elements of $\mathbf{x}_i$, and let S^p similarly be the submatrix of S^* consisting of rows with indices in p. Then, each row of $\boldsymbol{B}^*$ is given by

$$\mathbf{b}_i = (S^p)^{\dagger}(\mathbf{x}_i^p)^T$$

where $(S^p)^{\dagger}$ is the pseudo-inverse of S^p. With the new estimate of $\boldsymbol{B}^*$ at hand, let $\mathbf{x}_j$ be the jth column of X, let p be the indices of non-missing elements in $\mathbf{x}_j$, and let $\mathbf{x}_j^p$ be the column obtained by taking only the non-missing elements of $\mathbf{x}_j$. Let B^p similarly be the submatrix of B^* consisting of rows with indices in p. Then, the jth light source direction is given by,

$$\mathbf{s}_j = (B^p)^{\dagger}(\mathbf{x}_j^p)$$

After the new set of light sources S* has been calculated, the last two steps can be repeated until the estimate of B^* converges. The algorithm is very well behaved, converging to the global minimum within 10-15 iterations. Though it is possible to converge to a local minimum, we never observed this in simulation or in practice.

3.2 Enforcing Integrability

To predict cast shadows, we must reconstruct a surface and to do this, the vector field B^* must correspond to an integrable normal field. Since no method has been developed to enforce integrability during the estimation of B^*, *we* enforce it afterwards. That is, given B^* computed as described above, we estimate a matrix A $\in GL(3)$ such that B^*A corresponds to an integrable normal field; the development follows [17].

Consider a continuous surface defined as the graph of $z(x, \mathrm{y})$, and let b be the corresponding normal field scaled by an albedo (scalar) field. The integrability constraint for a surface is $z_{xy} = z_{yx}$ where subscripts denote partial derivatives. In turn, $\mathbf{b}$ must satisfy:

$$\left(\frac{b_1}{b_3}\right)_y = \left(\frac{b_2}{b_3}\right)_x$$

To estimate A such that $\mathbf{b}^{T}(x, \mathbf{yj} \doteq \mathbf{b}^{*T}$ **(T.** $y)A$, we expand this out. Letting the columns of A be denoted by A_1, A_2, A_3 yields

$$(\mathbf{b}^{*T}A_3)(\mathbf{b}_x^{*T}A_2) - (\mathbf{b}^{*T}A_2)(\mathbf{b}_x^{*T}A_3) =$$
$$(\mathbf{b}^{*T}A_3)(\mathbf{b}_y^{*T}A_1) - (\mathbf{b}^{*T}A_1)(\mathbf{b}_y^{*T}A_3)$$

which can be expressed as

$$\mathbf{b}^{*T}S_1\mathbf{b}_x^* = \mathbf{b}^{*T}S_2\mathbf{b}_y^* \tag{5}$$

where $S_1 = A_3A_2^T - A_2A_3^T$ and $S_2 = A_3A_1^T - A_1A_3^T$.

S_1 and S_2 are skew-symmetric matrices and have three degrees of freedom. Equation 5 is linear in the six elements of S_1 and S_2. From the estimate of $\boldsymbol{B}^*$ obtained using the method in Section 3.1, discrete approximations of the partial derivatives ($\mathbf{b}_x^*$ and $\mathbf{b}_y^*$) are computed, and then SVD is used to solve for the six elements of S_1 and S_2. In [17], it was shown that the elements of S_1 and S_2 are cofactors of A, and a simple method for computing A from the cofactors was presented. This procedure only determines six degrees of freedom of A. The other three correspond to the generalized bas relief (GBR) transformation [1] and can be chosen arbitrarily since GBR preserves integrability. The surface corresponding to B^*A differs from the true surface by GBR, i.e., $z^*(x,y) = \lambda z(x,y) + \mu x + \nu y$ for arbitrary λ, μ, ν with $\lambda \neq 0$.

3.3 Generating a GBR surface

The preceeding sections give a method for estimating the matrix B^* and then enforcing integrability; we now reconstruct the corresponding surface $\bar{z}(x,y)$. Note that $\bar{z}(x,y)$ is not a Euclidean reconstruction of the face, but a representative element of the orbit under a GBR transformation. Recall that both shading and shadowing will be correct for images synthesized from a transformed surface.

To find $\bar{z}(x,y)$, we use the variational approach presented in [7]. Then, it is a simple matter to construct an illumination cone representation that incorporates cast shadows. Using ray-tracing techniques for a given light source direction, we can determine the cast shadow regions and correct the extreme rays of C^*.

Figure 2 demonstrates the process of constructing the cone C^*. Figure 2.a shows the training images for one individual in the database. Figure 2.b shows the columns of the matrix B^*. Figure 2.c shows the reconstruction of the surface up to a GBR transformation. The left column of Fig. 2.d shows sample images in the database; the middle column shows the closest image in the illumination cone without cast shadows; and the right column shows the closest, image in the illumination cone with cast shadows.

4 Recognition

The cone C^* can be used in a natural way for face recognition, and in experiments described below, we compare three recognition algorithms to the proposed method. From a set of face images labeled with the person's identity (the *learning set)* and an unlabeled set of face images from the same group of people (the test set), each algorithm is used to identify the person in the test images. For more details of the comparison algorithms, see [2]. We assume that the face has been located and aligned within the image.

The simplest recognition scheme is a nearest neighbor classifier in the image space [4]. An image in the test set is recognized (classified) by assigning to it the label of the closest point in the learning set, where distances are measured in the image space. If all of the images are normalized to have zero mean and unit variance, this procedure is equivalent to choosing the image in the learning set that best *correlates* with the test image. Because of the normalization process, the result is independent of light source intensity.

As correlation methods are computationally expensive and require great amounts of storage, it is natural to pursue dimensionality reduction schemes. A technique now commonly used in computer vision - particularly in face recognition - is principal components analysis (PCA) which is popularly known as *Eigenfaces* [5, 11, 9, 16]. Given a collection of training images $\mathbf{x}_i \in \mathbb{R}^n$, a linear projection of each image $\mathbf{y}_i = W\mathbf{x}_i$ to an f-dimensional feature space is performed. A face in a test image x is recognized by projecting x into the feature space, and nearest neighbor classification is performed in $\mathbb{R}^f$. The projection matrix W is chosen to maximize the scatter of all pro-

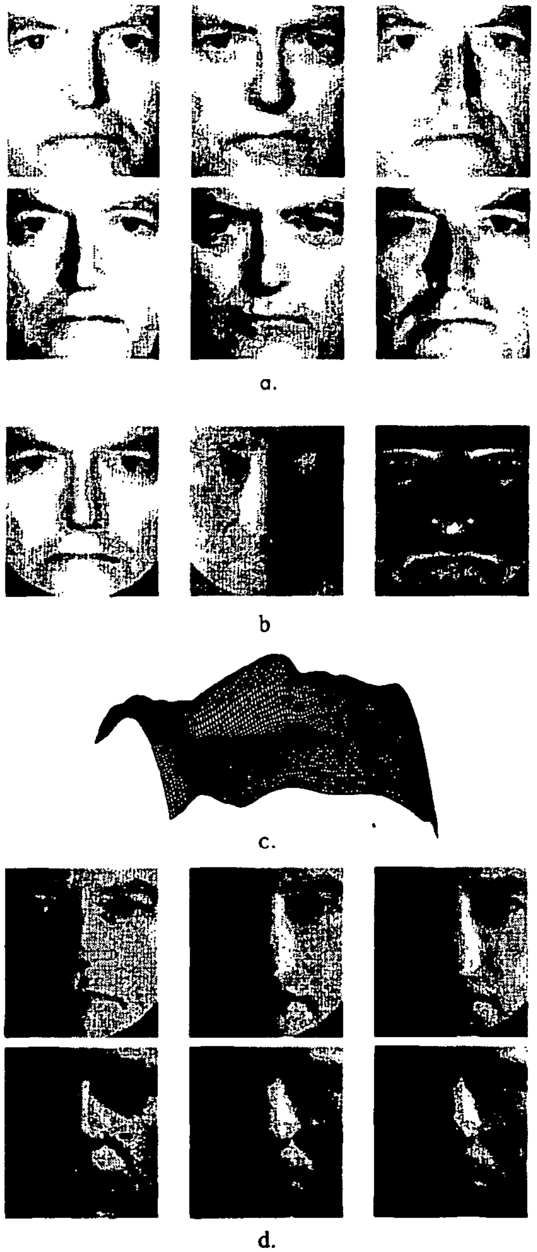


Figure 2: The figure demonstrates the process of constructing the cone C^*. a) the training images. b) matrix B^*. c) reconstruction up to a GBR transformation. d) sample images from database (left column); closest image in illumination cone without cast shadows (middle column); and closest image in illumination cone with cast shadows (right column).

jected samples. It has been shown that when f equals the number of training images, the Eigenface and Correlation methods are equivalent (See [2,11]). One proposed method for handling illumination variation in PCA is to discard from W the three most significant principal components; in practice, this yields better recognition performance [2].

A third approach is to model the illumination variation of each face as a three-dimensional linear subspace C as described in Section 2. To perform recognition, we simply compute the distance of the test image to each linear subspace and choose the face corresponding to the shortest distance. We call this recognition scheme the Linear *Subspace* method [1]; it is a variant of the photometric alignment method proposed in [13] and is related to [6,12]. While this models the variation in intensity when the surface is completely illuminated, it does not model shadowing.

Finally, given a test image $\mathbf{x}$, recognition using *ilZumination cones* is performed by first computing the distance of the test, image to each cone, and then choosing the face that corresponds to the shortest distance. Since each cone is convex, the distance can be found by solving a convex optimization problem. In particular, the non-negative linear least squares technique contained in Matlab was used in our implementation, and this algorithm has computational complexity $O(ne^2)$ where n is the number of pixels and e is the number of extreme rays.

5 Experimental Results

To test the effectiveness of these recognition algorithms, we performed a series of experiments on a database from the Harvard Robotics Laboratory in which lighting had been systematically varied [5,6]. In each image in this database, a subject held his/her head steady while being illuminated by a dominant light, source. The space of light source directions, which can be parameterized by spherical angles, was then sampled in 15” increments. See Figure 3. From this database, we used 660 images of 10 people (66 of each). We extracted five subsets to quantify the effects of varying lighting. Sample images from each subset are shown in Fig. 1. Subset 1 (respectively 2, 3, 4. 5) contains 30 (respectively 90, 130. 170, 210) images for which both the longitudinal and latitudinal angles of light source direction are within 15” (respectively 30°, 45°, 60°. 75”) of the camera axis.

All of the images were cropped (96 by 84 pixels) within the face so that the contour of the head was excluded. For the Eigenface and correlation tests, the images were normalized to have zero mean and unit variance, as this improved the performance of these methods. For the Eigenface method, we used twenty principal components - recall that performance approaches correlation as the dimension of the feature space is increased [2,11]. Since the first three principal components are primarily due to lighting variation and since recognition rates can be improved by eliminating them, error rates are also presented when principal components four through twenty-three are used. For the cone experiments, we tested two variations: in the first variation (Cones-attached), the representation was constructed ignoring cast shadows, and so extreme rays were generated directly from Eq. 3. In the second variation (Cones-cast), the representation was constructed as described in Section 3. In both variations, recognition was performed by choosing the face corresponding to the smallest computed distance to cone.

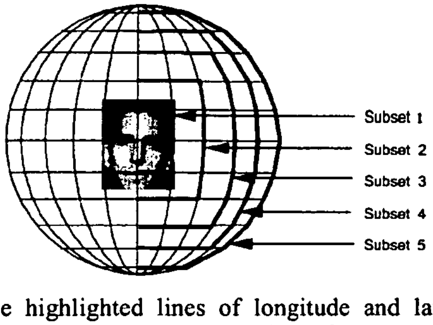


Figure 3: The highlighted lines of longitude and latitude indicate the light source directions for Subsets 1 through 5. Each intersection of a longitudinal and latitudinal line on the right side of the illustration has a corresponding image in the database.

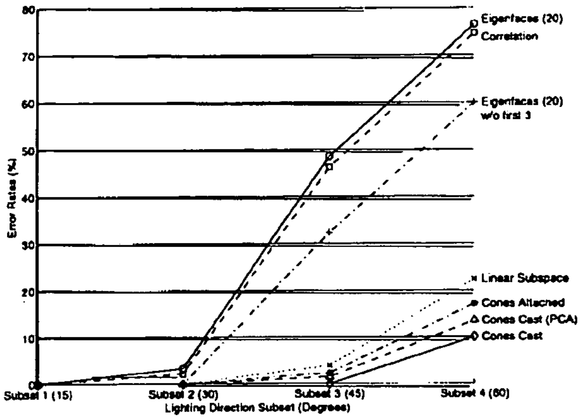
In our quest to speed up the recognition process using cones, we also employed principal components analysis (PCA). The collection of *all* images in the cones (with cast, shadows) is projeckd down to a 100-dimensional feature space. This is achieved by performing a linear projection of the form $y_i = W\mathbf{x}_i$, where the projection matrix W is chosen to maximize the scatter of all projected samples. A face in an image: normalized to have zero mean and unit variance, is recqgnized bv first projecting the image down to this 100-dimensional feature space and then performing nearest neighbor classification.

Mirroring the extrapolation experiment described in [2], each method was trained on samples from Subset 1 and then tested using samples from Subsets 2, 3, 4 and 5. (Note that when tested on Subset 1, all methods performed without error). Figure 4 shows the result from this experiment.

6 Discussion

From the results of this experiment, we draw the following conclusions:

- The illumination cone representation outperforms all of the other techniques.
- When cast shadows are included in the illumination cone, error rates are improved.
- PCA of cones with cast shadows outperforms all of the other methods except distance to cones with cast shadows. The small degradation in error rates is more than offset by the considerable speed up of more than one order of magnitude.



EXTRAPOLATING FROM SUBSET 1				
Method	Error Rate (%)			
	Subset 2	Subset 3	Subset 4	Subset 5
Correlation	2.2	46.2	74.7	86.6
Eigenface	3.3	48.5	76.5	86.6
Eigenface w/o 1st 3	0.0	32.3	60.0	80.6
Linear subspace	0.0	3.9	22.4	50.8
Cones-attached	0.0	2.3	17.1	43.8
Cones-cast (PCA)	0.0	1.5	13.5	39.8
Cones-cast	0.0	0.0	10.0	37.3

Figure 4: **Extrapolation:** When each of the methods is trained on images with near frontal illumination (Subset 1), the graph and corresponding table show the relative performance under more extreme light source conditions.

- For very extreme illumination (Subset 5), the Correlation and Eigenface methods completely break down, and exhibit results that arc slightly better than chance (90% error rate). The cone method performs significantly better, but certainly not well enough to be usable in practice. At this point, more experimentation is required to determine if recognition rates can be improved by either using more sampled extreme rays or by improving the image formation model.

The experiment described above was limited to the available dataset from the Harvard Robotics Laboratory. To perform more·extensive experimentation, we are constructing a geodesic lighting rig that supports 64 computer controlled xenon strobes. Using this rig, we will be able to modify the illumination at frame rates and gather an extensive image database covering a broader range of lighting conditions including multiple sources. The speed of acquisition will also permit us to readily obtain images of a large number of individuals. We will then perform more extensive experimentation with this newly gathered database.

References

[1] P. Belhumeur, D. Kriegman, and A. Yuille. The bas-relief ambiguity. In *Proc. IEEE* Conj. on Comp. *Vision* and *Patt. Recog.*, pages 1040-1046, 1997.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans.* Pattern *Anal.* Mach. *Intelligence*, 19(7):711 720, 1997. Special Issue on Face Recognition.

[3] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible lighting conditions. In *Proc. IEEE Conj. on Comp.* Vision and *Patt. Recog.*, pages 270-277, 1996.

[4] R. Brunelli and T. Poggio. Face recognition: Features vs templates. *IEEE* Trans. Pattern Anal. Mach. *Intelligence*, 15(10):1042-1053, 1993.

[5] P. Hallinan. A low-dimensional representation of human faces for arbitrary lighting conditions. In *Proc. IEEE* Conj. on *Comp. Vision* and *Putt. Recog.*, pages 995 999, 1994.

[6] P. Hallinan. *A Deformable Model for Face Recognition* Under *Arbitrary Lighting Conditions.* PhD thesis, Harvard University, 1995.

[7] B. Horn and M. Brooks. The variational approach to shape from shading. Computer *Vision, Graphics* and Image Processing, 35:174-208, 1992.

[8] D. Jacobs. Linear fitting with missing data: Applications to structure-from-motion and to characterizing intensity images. In *CVPR97*, pages 206-212, 1997.

[9] L. Sirovitch and M. Kirbv. Low-dimensional procedure for the characterization of human faces. J. *Optical Soc. of* America *A*, 2:519-524, 1987.

[10] Y. Moses, Y. Adini, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. In European Conj. on Computer *Vision*, pages 286- 296, 1994.

[11] H. Murase and S. Nayar. visual learning and recognition of 3-D objects from appearence. *Int. J. Computer Vision*, 14(5-24), 1995.

[12] S. Nayar and H. Murase. Dimensionality of illumination in appearance matching. *IEEE* Conj. on *Robotics* and *Automation*, 1996.

[13] A. Shashua. On photometric issues to feature-based object recognition. *Int. J. Computer Vision*, 21:99-122, 1997.

[14] H. Shum, K. Ikeuchi, and R. Reddy. Principal component analysis with missing data and its application to polyhedral object modeling. *PAMI*, 17(9):854-867, September 1995.

[15] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision, 9(2):134-154, 1992.

[16] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of* Cognztive Neuroscience, 3(1), 1991.

[17] A. Yuille and D. Snow. Shape and albedo from multiple images using integrability. In *Proc. IEEE* Conj. on Comp. Vision and *Patt. Recog.*, pages 158-164, 1997.

IEEE HOME | SEARCH IEEE | SHOP | WEB ACCOUNT | CONTACT IEEE

Membership Publications/Services Standards Conferences Careers/Jobs

IEEE Xplore® RELEASE 1.8

Welcome
United States Patent and Trademark Office

Help FAQ Terms IEEE Peer Review

Quick Links

» ABS

Welcome to IEEE Xplore®

- ○ Home
- ○ What Can I Access?
- ○ Log-out

Tables of Contents

- ○ Journals & Magazines
- ○ Conference Proceedings
- ○ Standards

Search

- ○ By Author
- ○ Basic
- ○ Advanced

Member Services

- ○ Join IEEE
- ○ Establish IEEE Web Account
- ○ Access the IEEE Member Digital Library

IEEE Enterprise

- ○ Access the IEEE Enterprise File Cabinet

Print Format

Search Results [PDF FULL-TEXT 320 KB] PREV NEXT DOWNLOAD CITATION

Using statistical models to interpret complex and v images

Taylor, C.J. Cootes, T.F. Edwards, G.
Dept. of Imaging Sci. & Biomed. Eng., Manchester Univ.

This paper appears in: **Applied Statistical Pattern Recognition (Ref. No. 1999/063), IEE Colloquium on**

Meeting Date: 04/20/1999
Publication Date: 20 April 1999
Location: Brimingham UK
On page(s): 6/1 - 6/4
Reference Cited: 7
Number of Pages: 106
Inspec Accession Number: 6276536

Abstract:
Model-based vision has been applied successfully to images of man-made obje proved much more difficult to develop model-based approaches to interpreting complex and variable structures such as faces or the internal organs of the hu The key problem is that of variability. Recent developments have shown that s patterns of variability in shape and grey-level appearance can be captured by models that can be used directly in image interpretation. The details of the ap outlined and practical examples from medical image interpretation and face re are used to illustrate how previously intractable problems can now be tackled successfully

Index Terms:
complex image interpretation computer vision face recognition grey-level appearance image processing medical images model-based vision shape variability statistical an statistical models variable structure images complex image interpretation computer v recognition grey-level appearance medical image processing medical images mode vision shape variability statistical analysis statistical models variable structure image

Documents that cite this document
There are no citing documents available in IEEE Xplore at this time.

Search Results [PDF FULL-TEXT 320 KB] PREV NEXT DOWNLOAD CITATION

Home | Log-out | Journals | Conference Proceedings | Standards | Search by Author | Basic Search | Advanced Search | Join IEEE | Web Account |

Using Statistical Models to Interpret Complex and Variable Images

C J Taylor, T F Cootes and G Edwards
Imaging Science and Biomedical Engineering, University of Manchester

Abstract

The ultimate goal of machine vision is image understanding – the ability not only to recover image structure but also to know what it represents. By definition, this involves the use of models that describe and label the expected structure of the world. Over the past decade, model-based vision has been applied successfully to images of man-made objects. It has proved much more difficult to develop model-based approaches to interpreting images of complex and variable structures such as faces or the internal organs of the human body (as visualised in medical images). In such cases it has even been problematic to recover image structure reliably, without a model to organise the often noisy and incomplete image evidence. The key problem is that of variability. To be useful, a model needs to be specific – that is, to be capable of representing only 'legal' examples of the modelled object(s). It has proved difficult to achieve this whilst allowing for natural variability. Recent developments have overcome this problem; it has been shown that specific patterns of variability in shape and grey-level appearance can be captured by statistical models that can be used directly in image interpretation. The details of the approach are outlined and practical examples from medical image interpretation and face recognition will used in the presentation to illustrate how previously intractable problems can now be tackled successfully.

1. Introduction

The majority of tasks to which machine vision might usefully be applied are 'hard'. The examples we use in this paper are from medical image interpretation and face recognition, though the same considerations apply to many other domains.

The most obvious reason for the degree of difficulty is that most non-trivial applications involve the need for an automated system to 'understand' the images with which it is presented – that is, to recover image structure *and* know what it means. This necessarily involves the use of models that describe and label the expected structure of the world. Real applications are also typically characterised by the need to deal with complex and variable structure – faces are a good example – or with images that provide noisy and possibly incomplete evidence – medical images are a good example, where it is often impossible to interpret a given image without prior knowledge of anatomy.

Model-based methods offer potential solutions to all these difficulties. Prior knowledge of the problem can, in principle, be used to resolve the potential confusion caused by structural complexity, provide tolerance to noisy or missing data, and provide a means of labelling the recovered structures. We would like to apply knowledge of the expected shapes of structures, their spatial relationships, and their grey-level appearance to restrict our automated system to 'plausible' interpretations.

Of particular interest are generative models – that is, models sufficiently complete that they are able to generate realistic images of target objects. An example would be a face model capable of generating convincing images of any individual, changing their expression and so on. Using such a model, image interpretation can be formulated as a matching problem: given an image to interpret, structures can be located and labelled by adjusting the model's parameters in such a way that it generates an 'imagined image' which is as similar as possible to the real thing.

Because real applications often involve dealing with classes of objects that are not identical – for example faces – we need to deal with variability. This naturally leads to the idea of deformable models – models that maintain the essential characteristics of the class of objects they represent, but which can deform to fit a range of examples. There are two main characteristics we would like such models to possess. First, they should be *general* – that is, they should be capable of generating any plausible example of the class they represent. Second, and crucially, they should be *specific* – that is, they should *only* be capable of generating 'legal' examples – because, as we noted earlier, the whole point of using a model-based approach is to limit the attention of our system to plausible interpretations. In order to obtain specific models of variable objects, we need to acquire knowledge of how they vary. Although it is possible, for some classes of object, to predict the forms of variability that will be encountered, generally, the only realistic approach is to learn this from a set of examples.

In the remainder of this paper we outline our approach to modelling shapes, spatial relationships and grey-level appearance, and show how these models can be used in image interpretation.

2. Modelling Shapes and Spatial Relationships

Our approach to modelling shapes and spatial relationships has been described previously [1, 2]. The first step is to extract a vector representation of each example shape in a training set of similar shapes. We describe the shapes using a set of landmark points placed at similar positions on each example. We can achieve a consistent representation by choosing a set of primary landmark points at well-defined points, then adding equally spaced points to represent the rest of the boundary. For example, we might describe a set of hand shapes in a consistent way by placing landmarks at the tips of the fingers and the cracks between them. If there are *n* points, the result is a vector $\mathbf{x}_i = \{x_{i1}, y_{i1}, x_{i2}, y_{i2} \cdots x_{in}, y_{in}\}$ containing 2*n* ordinates representing each example *i*. In order to standardise this representation it is necessary to align the set of examples into a common co-ordinate frame using, for example, a Procrustes analysis.

The next step is to think about the set of training examples in the vector space defined by $\mathbf{x}$. Typically the training set will form a cloud in a very high dimensional space. An important observation is that the values of different components of the vector will tend to be correlated, and that it is these correlations which tell us about the invariant properties of the class of shapes. For example, if we consider two components of the vector, $x_1$and x_2, as the horizontal positions of two points on the same edge of a finger, they will tend to move together as the hand changes shape. As a result, we find that, generally, the subspace in which 'legal' examples of a class of shapes are found has much lower dimensionality than the shape space in which it is embedded. The major axes of this subspace can be found by principal component analysis (or non-linear equivalents [3, 4]) and results in a linear model which is able to reconstruct any of the examples in the training set.

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s \tag{1}$$

A given example $\mathbf{x}$ can be reconstructed from a weighted sum of the mean shape $\bar{\mathbf{x}}$ and a set of linearly independent *modes of variation*, $\mathbf{P}_s$. These modes are the eigenvectors of the covariance matrix of the set of training shapes; the weight vector $\mathbf{b}_s$ is a set of shape parameters which, given the model, provide a unique description of the example shape. Examples of 'legal' shapes can be synthesised by choosing vales for the $\mathbf{b}_s$parameters within limits determined from the training set. Since the eigenvectors which form $\mathbf{P}_s$ are orthogonal, equation 1 can be solved easily for the shape parameters, given an example shape.

$$\mathbf{b}_s = \mathbf{P}_s^T (\mathbf{x}_i - \bar{\mathbf{x}}) \tag{2}$$

Although our illustrative example of a hand shape has a simple closed boundary, it is important to note that, since the basis for our representation is simply a set of points, complex, multi-part objects can also be modelled, allowing shapes and spatial relationships to be treated in a unified manner. Examples will be given in the presentation.

3. Modelling Local Grey-Level Appearance

In parallel with building a model of shape and spatial relationships, we build a statistical model of the grey-level pattern in the vicinity of each model point. These local grey-level models are typically chosen to represent the grey-level appearance along linear profiles sampled at the model points, perpendicular to the model boundary, and take the form of factor models. These models are important in image search and allow a matching score to be defined between any image patch and the expected grey-level pattern at a given model point. The details have been presented previously [1].

4. Interpretation – Active Shape Models

So far we have seen how we can build statistical models of shapes, spatial relationships and local grey-level appearance. In this section we show how these models can be used in automatic image interpretation. The *Active Shape Model* (ASM) approach described here has been described in more detail previously [1]. It has much in common with the 'snakes' or 'active contours' of Kass and Witkin, but with the crucial difference that we use the model to apply global constraints to shapes and spatial relationships. The idea is to place an initial model instance into the image and to refine it iteratively. Each model point tries to move towards the appropriate image feature by finding a point close to its current position at which there is a better match to its local grey-level model.

A search profile is set up at each model point, normal to the current model contour. At some position along each profile – hopefully at the true boundary contour of the object – we find a better match to the local grey-level model. The key step is that we try to move towards these better matches by updating the parameters of the model– its position, scale, orientation and shape – not the positions of the points directly. This involves two steps: first we cast the proposed shape into the model frame by finding the translation, orientation and scale which align it as closely as possible to the current model; we then compute new shape parameters using equation 2, impose limits on their values to ensure a plausible shape, and project back into the image using equation 1. This ensures that our new estimate is always a 'legal' solution, because the model is only capable of generating legal solutions.

In practice, the speed and robustness of ASM search can be improved significantly by using a multi-resolution approach [5]. During training, a Gaussian pyramid is constructed from each image, and local grey-level models are trained for each level of the pyramid. For new images, ASM search starts at the coarsest level of a similar image pyramid, using the corresponding grey-level model; the search profile length is chosen to allow model points to move some distance to their targets. When convergence is detected at the current scale, the next finer scale is selected, and model refinement continues from the existing solution using a shorter search profile. This is repeated until a solution has been found at the finest scale. Using the multi-scale approach, ASMs typically converge to the correct solution, even given a very poor initialisation.

5. Full Appearance Models

The ASM method uses only local models of grey-level appearance. In this section we describe the construction of full generative models of both shape and grey-level appearance of the entire object. A more complete description is given in [6]. Our approach is to combine a model of shape variation with a shape-normalised grey-level model. The corresponding sets of hand-placed landmark points are used in a warping algorithm; this deforms each training image such that the object assumes the same standard shape (the average shape over the training set). We then sample the grey-level information, $\mathbf{g}$, from the object in the *shape-normalised* image. By applying PCA to this data we obtain a linear model analogous to the shape model.

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \qquad (3)$$

The shape and appearance of any example can thus be summarised by the vectors $\mathbf{b}_s$ and $\mathbf{b}_g$. Since there may be correlations between shape and grey-level variation, we apply a further PCA to the data as follows. For each example we generate the concatenated vector

$$\mathbf{b} = \begin{pmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix} = \begin{pmatrix} \mathbf{W}_s \mathbf{P}_s^T (\mathbf{x} - \bar{\mathbf{x}} \\ \mathbf{P}_g (\mathbf{g} - \bar{\mathbf{g}}) \end{pmatrix} \qquad (4)$$

where $\mathbf{W}_s$ is a diagonal matrix of weights for each shape parameter, allowing for the difference in units between the shape and grey models. We apply a PCA to these vectors, yielding a combined model

$$\mathbf{b} = \mathbf{Q}\mathbf{c} \qquad (5)$$

where $\mathbf{Q}$ are the eigenvectors of $\mathbf{b}$, and $\mathbf{c}$ is a vector of *appearance* parameters controlling both the shape and grey-levels appearance of the model. An example image can be synthesised for a given $\mathbf{c}$ by generating the shape-free grey-level image $\mathbf{g}$ and warping it using the control points described by $\mathbf{x}$.

6. Interpretation – Active Appearance Models

To use the full appearance models described above for image interpretation, we need a method of automatically matching the models to image data. Here we describe an efficient algorithm for adjusting the model parameters to match the image. A full description is given in [7].

We seek the optimum set of model parameters (including location, orientation and scale) that best describe the image data. A suitable metric for the quality of the match between model and image is $\Delta = |\delta \mathbf{g}|^2$ where $\delta \mathbf{g}$ is the vector of differences between the grey-level values in the image and a corresponding instance of the model. We could seek to vary the model parameters to minimise Δ, treating this as a general optimisation problem. This is, however, not a practical approach given that a typical appearance model may have 100 or more parameters.

The Active Appearance Model algorithm uses the full vector $\delta \mathbf{g}$ to drive the search, rather than a scalar objective function. We note that each attempt to match the model to a new image is actually a similar optimisation problem and that solving a general optimisation problem from scratch is unnecessary. The AAM attempts to learn how to solve this class of problems in advance. By providing a-priori knowledge of how to adjust the model parameters during image search, we can obtain an efficient run-time algorithm. In particular, the AAM uses the spatial pattern in $\delta \mathbf{g}$, to encode information about how the model parameters should be changed in order to achieve a better fit.

For example, if the largest differences between a face model and a face image occurred at the sides of the face, this would imply that a parameter that modified the width of the model face should be adjusted.

The method works by learning from an annotated set of training example for which the 'true' model parameters are known. For each example in the training set, a number of known model displacements $\delta \mathbf{c}$ are applied, and the corresponding difference vector $\delta \mathbf{g}$ is recorded. Once enough training data has been generated, multivariate multiple regression is applied to learn a linear relationship between the model displacement and image difference. Image search then involves placing the model in the image and measuring the difference vector. The learnt regression model is used to predict a change in the model parameters likely to give a better match. The process is iterafed to convergence.

7. Summary

The methods outlined in the paper provide a principled and flexible basis for developing practical systems to solve difficult image interpretation problems - particularly in medical image analysis and face recognition, but also in other application domains. The presentation will be illustrated with examples.

[1] T. Cootes, A. Hill, and C. Taylor, "The use of Active shape models for locating structures in medical images.," *Image and Vision Computing*, vol. 12, pp. 355-366, 1994.

[2] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models their training and application," *Computer Vision and Image Understanding*, vol. 61, pp. 38-59, 1995.

[3] P. Sozou, T. Cootes, C. Taylor, and E. Di Mauro, "Non-linear generalisation of point distribution models using polynomial regression," *Image and Vision Computing*, vol. 13, pp. 451-457, 1995.

[4] P. Sozou, T. Cootes, C. Taylor, E. Di Mauro, and A. Lanitis, "Non-linear point distribution modelling using multi-layer perceptron," *Image and Vision Computing*, vol. 15, pp. 457-463, 1997.

[5] T. Cootes, C. Taylor, and A. Lanitis, "Active shape models: Evaluation of a multi-resolution method for improving image search.," Proceedings of *British Machine Vision Conference*: BMVA press, pp. 327-338, 1994.

[6] G. Edwards, A. Lanitis, C. Taylor, and T. Cootes, "Statistical models of face images: Recent advances," Proceedings of *British Machine Vision Conference*: BMVA Press, pp. 765-774, 1996.

[7] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," Proceedings of *ECCV*: Springer Verlag, pp. 484-498, 1998.

Welcome
United States Patent and Trademark Office

Welcome to IEEE Xplore®

○ Home
○ What Can I Access?
○ Log-out

Tables of Contents

○ Journals & Magazines
○ Conference Proceedings
○ Standards

Search

○ By Author
○ Basic
○ Advanced

Member Services

○ Join IEEE
○ Establish IEEE Web Account
○ Access the IEEE Member Digital Library

IEEE Enterprise

○ Access the IEEE Enterprise File Cabinet

Print Format

Search Results [PDF FULL-TEXT 804 KB] PREV NEXT DOWNLOAD CITATION

Request Permissions RIGHTSLINK

Modelling the variability in face images

Edwards, G.J. Lanitis, A. Taylor, C.J. Cootes, T.F.
Dept. of Med. Biophys., Manchester Univ., UK;

This paper appears in: **Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on**

Meeting Date: 10/14/1996 - 10/16/1996
Publication Date: 14-16 Oct. 1996
Location: Killington, VT USA
On page(s): 328 - 333
Reference Cited: 26
Number of Pages: xxvi+386
Inspec Accession Number: 5425676

Abstract:
Model based approaches to the interpretation of face images have proved very successful. We have previously described statistically based models of face sh grey-level appearance and shown how they can be used to perform various co interpretation tasks (Lanitis et al., 1995). In the paper we describe improved modelling, which couple shape and grey-level information more directly than o methods, isolate the changes in appearance due to different sources of variab (person, expression, pose, lighting) and deal with nonlinear shape variation. W that the new methods are better suited to interpretation and tracking tasks

Index Terms:
computational geometry face recognition image coding image recognition statistical expression face image recognition face image variability modelling face shape grey appearance image coding image interpretation lighting model based approaches n shape variation person pose statistical models tracking tasks

Documents that cite this document
There are no citing documents available in IEEE Xplore at this time.

Search Results [PDF FULL-TEXT 804 KB] PREV NEXT DOWNLOAD CITATION

Modelling the Variability in Face Images

G.J. Edwards, A. Lanitis, C.J. Taylor, T. F. Cootes
Department of Medical Biophysics, University of Manchester, UK
email: {gje,lan,ctaylor,bim}@sv1.smb.man.ac.uk

Abstract

Model based approaches to the interpretation of face images have proved very successful. We have previously described statistically based models of face shape and grey-level appearance and shown how they can be used to perform various coding and interpretation tasks. In the paper we describe improved methods of modelling which couple shape and grey-level information more directly than our existing methods, isolate the changes in appearance due to different sources of variability (person, expression, pose, lighting), and deal with non-linear shape variation. We show that the new methods are better suited to interpretation and tracking tasks.

1. Introduction

Model-based approaches to the interpretation and coding of face images have proved very successful. Methods described so far include: Modelling grey-level variation using eigenfaces [8,24], models based on class specific projections [1], combined shape and grey level models [5,20], models based on the physical and anatomical structure of faces [23], 3D models [15], hand- crafted shape models [26], local non-linear shape manifolds[2], and models based on elastic meshes coubled with local intensity pattern descriptions[11]. Compehensive literature reviews of these techniques and other techniques related to face interpretation can be found in [3,21,25].

The success of a model-based approach relies on the quality of the face model used. In general the models must fulfill two main criteria: generality and specificity. General models are those that account for all possible sources of appearance variation in face images. Specific models constrain the variability allowed so that only 'legal' examples can be generated. In addition to these criteria successful models should be compact and also have the potential to be used in image search algorithms. In the past we have achieved promising results using a model-based approach [12,13]. In this paper we describe further developments of our models of facial appearance; by using the improved models we aim to improve the performance of our system. In particular we describe how shape and global grey-level variation can be modeled using a single rather than separate models. We also describe how the different sources of variability can be isolated, given a suitable training set of images. Isolating the sources of variation can be useful in image synthesis and in tracking; where the dynamics of the different sources of variation will differ. The models we have previously described are based on a linear formulation. We present the results of shape modeling and image search experiments using a non-linear formulation for modeling shape. These show that more accurate results are obtained using the non-linear approach.

2. Overview of Our Previous Work

Our approach can be divided into two main phases: modeling, in which flexible models of facial appearance are generated, and interpretation, in which the models are used for coding and interpreting face images. Flexible models [7] are generated from a set of training examples, by statistical analysis. As a result of the analysis training examples can be reconstructed/ parameterized using:

$$\mathbf{X} = \mathbf{X}_m + \mathbf{Pb}$$

where $\mathbf{X}$ is a training example, $\mathbf{X}_m$ is the mean example, $\mathbf{P}$ is the matrix of eigenvectors and $\mathbf{b}$ is a vector of weights , or model parameters. Flexible models can be used for modeling shape and/or grey-level variation. We model the shapes of facial features and their spatial relationships using a single flexible shape model (a Point Distribution Model) [7]. The effect of the most significant shape parameters is shown in figure 1 (For the experiments described in this section the Manchester Face Database [17] was used). We have previously shown [6] that shape models of this form are more specific than other types of deformable models (e.g. FEM models [19]) and thus lead to more robust model-based interpretation results.

0-8186-7713-9/96 $5.00 © 1996 IEEE

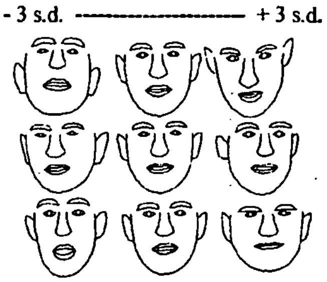


Fig 1. The first 3 modes of shape variation

Our shape model is augmented with flexible grey-level models using two complementary approaches. In the first we generate a flexible grey-level model of 'shape-free' appearance by deforming each face in the training set to have the same shape as the mean face (the effect of the main parameters of this model is shown in figure 2). In the second approach we use a large number of local profile models, one at each landmark point of the shape model. The first approach is more complete but the second is more robust to partial occlusion[12].

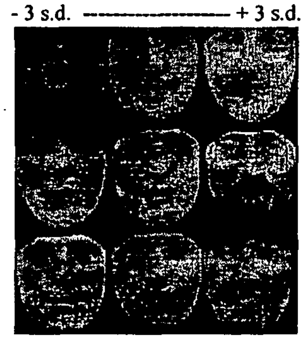


Fig 2. First 3 modes of shape-free grey-level variation

Shape and grey-level models are used together to describe the overall appearance of each face; collectively we refer to the model parameters as appearance parameters. It is important to note that the coding we achieve is reversible - a given face image can be reconstructed from its appearance parameters. When a new image is presented to our system, facial features are located automatically using Active Shape Model (ASM) search [4,7] based on the flexible shape model obtained during training. The resulting automatically located model points are transformed into shape model parameters. Grey-level information at each model point is collected and transformed to local grey-level model parameters. Then the face is deformed to the mean face shape and the grey-level appearance is transformed into the parameters of the shape-free grey-level model. We have presented [12,13] results showing that this representation can be used for image reconstruction, person identification (including gender recognition), expression recognition and pose recovery.

3. Training Combined Shape and Grey-Level Models

In our previous work we used separate shape and grey-level models to represent facial appearance. However, shape and grey-level variations may be correlated; certain combinations of shape and grey-level modes may correspond to illegal facial reconstructions, thus the overall model is not specific enough. For example the shape mode of variation responsible for opening and closing the mouth is correlated with the grey level mode responsible for the appearance of teeth. We have generated a combined shape and grey-level model in order to overcome this problem. We first train individual shape and shape-free grey-level models [12,13] and convert all training examples to the corresponding model parameters. This results in the representation of training examples by a vector containing both shape and grey-level parameters. Principal component analysis is applied to the new training vectors in order to extract the combined shape and grey-level modes of variation. Before applying the final PCA we scale the shape parameters so that their variance within the training set is equal to the variance of the grey-level parameters. Figure 3 shows the first few modes of the combined shape / grey-level model trained using images from the Home Office Database [9]. Figure 4 shows parametric reconstructions of original images using a combined shape and grey-level model. (This example included hair.)

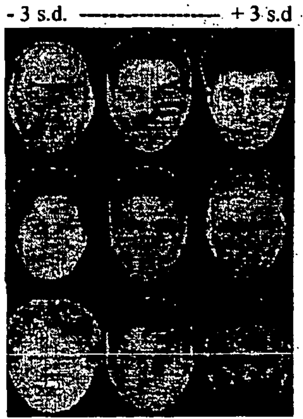


Fig 3. First three modes of combined model

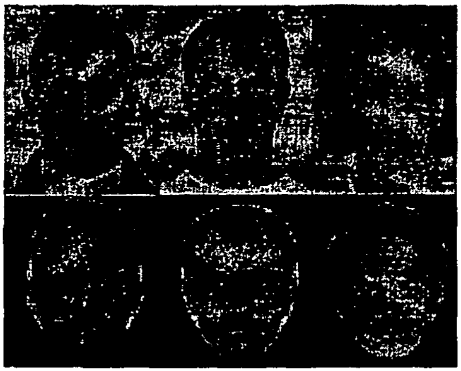


Fig 4. Original images with reconstructions

4. Isolating Sources of Variation

There are four main sources of appearance variation in face images:

- Pose changes.
- Lighting changes.
- Changes due to difference in individual appearance.
- Changes due to expression, or other face movement, e.g. speaking.

In the models we have presented so far, individual modes of variation tend to be associated with a particular source of variation. However, this is not guaranteed to be the case. If modes corresponding to the different sources of variation could be isolated reliably there would be several benefits. Firstly, for image synthesis applications we could manipulate chosen characteristics without changing others, for example, expression without ID. Secondly, for tracking, we could model the variation of the different components independently, for example, ID modes would be expected to remain constant whilst others would vary over time. Finally we hope that visualization of the modes, for example, the expression modes, will provide insight into the factors involved in recognition.

4.1 Discriminant Analysis

To achieve the desired isolation of the variation arising from the different sources, we employ canonical discriminant analysis over the discrete range of classes of interest. The classification is clear for person ID, where a class corresponds to a particular individual. For expression, we choose a classification based on seven expressions. The models shown in this paper were trained using the pooled results from experiments involving 30 observers assigning one of seven expressions to each image (These experiments used the Manchester Expression Database[16]).

The goal of canonical discriminant analysis is to define linear combinations of a set of variables, which separate the classes as well as possible. If there are p variables, in a vector **X**, the ith discriminant function Z_i is given by:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$$

Finding the coefficients a_{ij} is an eigenvalue problem. The within class covariance matrix **W**, and the total sample covariance matrix **T** are found, and from these the between class covariance is computed(see [18]):

$$\mathbf{B} = \mathbf{T} - \mathbf{W}$$

The discriminant functions are the eigenvectors of the matrix $\mathbf{W}^{-1}\mathbf{B}$, with the corresponding eigenvalues describing the amount of separation, the first function reflecting as much class difference as possible, and so on. For computational simplicity, we perform this analysis on the **b**-vectors for shape and grey-level appearance, obtained using our conventional principle component analysis. The **b**-vector for a particular example is given by:

$$\mathbf{b} = \mathbf{D}\mathbf{d}$$

where **d** is a vector of discriminant parameter weights and **D** is the matrix of unit eigenvectors defining the discriminant functions. The original shape/grey-level vector is therefore given by:

$$\mathbf{X} = \mathbf{X}_m + \mathbf{P}\mathbf{D}\mathbf{b}$$

The maximum rank of **D** is (no. classes - 1) which is normally less than the number of b -values. Examples can be parameterized and reconstructed using the coefficients of the d-vector (which we call *Discriminant Model Parameters.)* Reconstructions of three canonical discriminant modes for expression are shown in figure 5. Figure 6 shows the equivalent modes for changes between individual (ID modes).

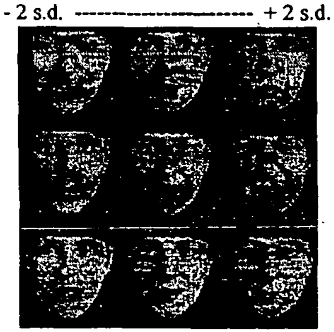


Fig 5. Three major discriminant modes for expression.

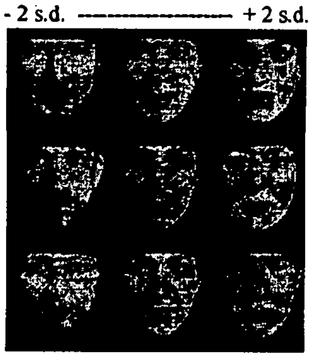


Fig 6. Three major discriminant modes for individual.

4.2 Removing Discriminant Modes

After finding discriminant modes, we project an example to the least squares approximation in discriminant space. The difference between the example and it's approximation is then used as a new training example. The new set of examples should contain no variation due to the sources defined by the discriminant vectors. A principle component analysis is performed on these examples, producing a model with only those modes of variation orthogonal to the discriminant space. Figure 7 shows the first three modes of the model after first removing variation due to change of individual.

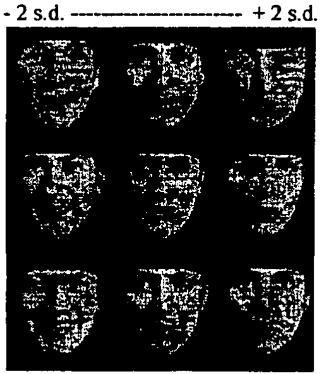


Fig 7. First 3 principle components after ID variation removed

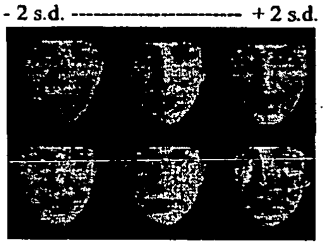


Fig 8. First 2 principle components after removing expression and ID

Encouragingly, there is only slight change in the perceived individual, which indicates a good degree of separability between individual and expression modes of variation. There remains, of course, variation due to pose and lighting conditions.

In figure 8 both expression and ID have been removed which ought to leave only variation due to pose and lighting. It appears to make very little difference in which order the two sources of variation are removed. The training set used in this example did not feature a great deal of pose variation; this accounts for the relatively small amount of variation shown in the figure. The idea that the different sources of variation are completely separable is, of course, a simplification. For example, different individuals have characteristic smiles. The results shown in figures 5,6,7 and 8 suggest that the assumption of separability is, however, a useful approximation.

5. Using Non-Linear Shape Models

The shape model described in our earlier work[12,13] is based on a linear formulation which may fail when we attempt to model extreme pose variation face images. In this section we describe how a non-linear shape model can be built using a Multi-Layer Perceptron (MLP) and describe experiments for assessing the goodness of the non-linear model in locating facial features when compared with the results obtained using the linear model.

5.1Non-Linear PDMs using Multi-Layer Perceptrons

The use of multilayer perceptrons (MLPs) for carrying out non-linear principal component analysis has been described by Kramer [10]. His approach involves training an MLP to give a set of outputs which are as close as possible to the inputs, over a training set of examples.

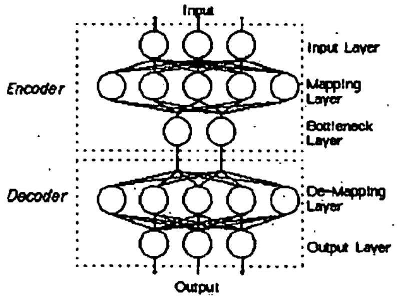


Fig 9. Structure of the MLP

Recently we have described [14,22] how non-linear PDMs can be formulated using a similar approach.

During the training procedure we perform an initial linear PCA on the training shape data. Using the basis functions calculated during this procedure we convert all training examples to principal components and feed them into an MLP of the form shown in figure 9.

During the training phase the weights of the network are adjusted so that the inputs of the network arc faithfully reconstructed at the output nodes. The key feature is a "bottleneck" layer with a small number of neurons. In order to achieve outputs equal to the inputs, the MLP is forced to code the data into a number of components equal to the number of neurons in the bottleneck layer thus effecting a non-linear dimension reduction. Once the MLP has been trained using the conjugate gradient decent algorithm, we split it into an encoder and a decoder. We use the encoder to obtain the coded representations of the training set and the decoder to rcconstruct training examples given the coded representation. Figure 10 shows schematically the parametrization and reconstruction of training shapes using the non-linear PDM. Non-linear PDMs can be used in image search in a similar scheme as a linear PDM [14]. A combined image search strategy which used the non-linear model in the initial stages of the search followed by refinement using the local model was also investigated[14].

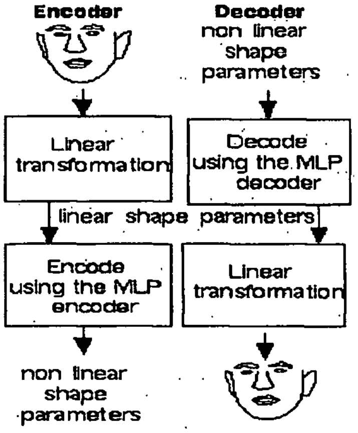


Fig 10. Shape Space / Parameter Space Projections

5.2 Performance of Non-Linear Face Models

We trained a linear PDM and a non-linear PDM (or MLP-PDM) using the samc training data as that used previously (see section 2). The linear PDM needed 16 shape parameters to explain 99% of the variability in the training set whilst the MLP-PDM needed only eight. This implies that there were non-linear dependencies between the modes of variation in the linear model and that, as a result, it must be capable of generating illegal solutions. The MLP-PDM model was substantially more compact and thus more specific. The first few significant modes of variation for both the linear PDM and MLP-PDM are similar.

Systematic experiments compared the performance of linear PDMs and MLP-PDMs in the context of their ability to locate facial features using ASM. We tested the fitting procedure by fitting the model to 40 face images, in two main experiments. For the first experiment the initial pose was chosen randomly within the following limits: rotation of +/-20 degrees, displacement from the correct position by +/- 30 pixels, and starting scale of 0.6 to 1.4 of the mean scale. These limits usually resulted in a very poor starting point for the iterative search procedure. For the second set of experiments the initial pose was defined within narrower limits: rotation of +/- 10 degrees, displacement from the correct position by +/- 10 pixels, and starting scale of 0.8 to 1.2 of the mean scale. For both experiments, the model was initialized to the mean shape. For each test image we fitted the models using three different initial poses giving a total of 120 number of trials. The correct positions for all 144 model points were marked manually on all thc test images. At each iteration of the ASM search the goodness of fit, defined as the mean Euclidean distance, d, between the positions of thc model points and their correct positions, was calculated. Figure 11 summarizes the results of the experiments; the graphs show the average value of d against the iteration number, over all 120 model fitting trials. In experiment 1 ASM search using an MLP-PDM performs better than the linear PDM. For experiment 2 where the starting position of the model is on avcrage closer to the target, image search using a linear PDM performs better. The performance of the combined method is better in both experiments than search using either a lincar PDM or an MLP-PDM alone.

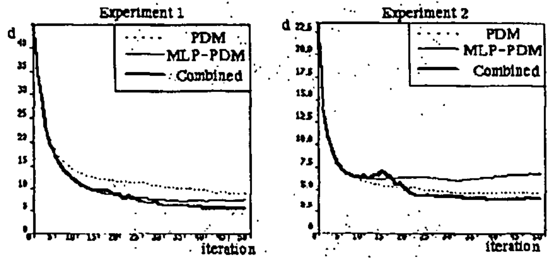


Fig 11. 2. Results for the image search experiments

6. Conclusions

We have described our work in progress on developing improved models of facial appcarance. We have shown that it is feasible to model both grey-level and shape variation using a single model resulting in a more specific overall model. By using discriminant analysis techniques we have shown that modes corresponding to different sources of variation can be isolated. The next step is to use these decoupled modes to track faces in image sequences. We have also described how a non-linear PDM can be built and we have presented results for locating facial features. These results show that, when the initial placement of the model is bad, image search using the non-linear shape model performs better than image search using a linear PDM. The reason for this is the ability of the MLP-PDM to be more specific to the class exemplified by the training set. The domain of possible solutions is reduced since only plausible solutions are allowed, resulting in an increased chance of locating image objects successfully. However, when the starting position is good, image search using a linear PDM performs better, since in this case models are unlikely to be driven to illegal shapes. Image search using a combination of linear and non-linear models proved to be the most robust and accurate in our experiments.

7. Acknowledgments

We would like to thank EPSRC and British Telecom for supporting this research. Also, the Home Office (UK) who provided the Home Office Database, Dr Jane Whittaker from Royal Manchester Children's Hospital and The Manchester Actors Center for providing the Manchester Expression Database, the volunteers who gave face images for the Manchester Face Database, Mr John Connell and Mr Ian Wilks.

8. References

[1] P. Belhumeur, J.Hespanha, D.Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. Procs ECCV96 (eds. B.Buxton and R.Cipolla), 1, pp 45-58, Springer, 1996.

[2] C. Bregler, S.Omohundro. Nonlinear Manifold Learning for Visual Speech Recognition. Procs. of the 5th International Conference on Computer Vision, pp 494-499, IEEE Computer Society Press, Cambridge, USA, 1995.

[3] R. Chellapa, C.L. Wilson and S. Sirohey. Human and machine Recognition of Faces: A Survey. Procs of the IEEE, vol 83, no 5, 1995.

[4] T.F. Cootes, C.J. Taylor and A. Lanitis. Active Shape Models: Evaluation of a Multi-Resolution Method For Improving Image Search. Procs. of the 5th British Machine Vision Conference, vol 1, pp 327-336, (ed. Edwin Hancock), BMVA Press, 1994.

[5] T.F.Cootes and C.J.Taylor. Modeling Object Appearance Using the Grey - Level Surface, Procs. of the 5th British Machine Vision Conference (ed E. Hancock), vol 2, pp 479-488, BMVA Press, 1994.

[6] T.F.Cootes and C.J. Taylor. Combining Point Distribution Models With Models Based on Finite Element Analysis. Image and Vision Computing, Vol 13, no 5, pp 403-409, 1995.

[7] T.F. Cootes, C.J. Taylor, D.H. Cooper and J. Graham. Active Shape Models - Their Training And Application. Computer Vision Graphics and Image Understanding, Vol 61, no 1, pp 38-59, 1995.

[8] I. Craw and P. Cameron. Face Recognition by Computer. Procs of British Machine Vision Conference 1992, pp 489-507, eds. David Hogg and Roger Boyle, Springer Verlag, 1992.

[9] The Home Office Database contains about 500 images, one per individual. Most of the variation in the database is due to inter-individual appearance. The individuals in the database cover a wide age range (20 -60 years), both genders and different ethnic origins.

[10] M.A. Kramer, Nonlinear Principal Component Analysis Using Autoassociative Neural Networks, AIChE Journal pp 233-243, 1991.

[11] M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C. Malsburg, R.P. Wurtz and W. Konen. Distortion Invariant Object Recognition in the Dynamic Link Architecture. IEEE Transactions on Computers, Vol 42, no 3, pp 300-311, 1993.

[12] A.Lanitis, C.J. Taylor, T.F.Cootes. Automatic Face Indentification System using Flexible Appearance Models, Image and Vision Computing, Vol 13, no 5, pp 393-401, 1995.

[13] A.Lanitis, C.J. Taylor, T.F.Cootes. A Unified Approach To Coding and Interpreting Face Images. Procs. of the 5th International Conference on Computer Vision, pp 368-373, IEEE Computer Society Press, Cambridge, USA, 1995.

[14] A.Lanitis, P.D.Sozou, C.J.Taylor, T.F.Cootes and E.C. Di-Mauro. A General Non-Linear Method For Modelling Shape Variation and Locating Image Objects. To Appear in the Procs of the International Conference of Pattern Recognition, 1996.

[15] H. Li, P. Roivainen, R. Forchheimer. 3D Motion Estimation in Model-Based Facial Coding. IEEE Trans. of PAMI, vol 15, no 6, pp 545-555, 1993.

[16] The Manchester Expression Database contains 400 images (about 22 images from 16 individuals). All 16 individuals were actors who posed realistic facial expressions while listening to descriptions of situations, used to assist their acting. The database contains appearance variations due to individual appearance, lighting, expression and pose.

[17] The Manchester Face Database contains 23 images of each of 30 individuals. (10 training, 10 test and three difficult test images per individual). Significant individual variation, pose, expression, and lighting variation exist in this database. The database is publicly available at: http://peipa.essex.ac.uk/ftp/ipa/pix/faces/manchester

[18] B.J.F. Manly, *Multivariate Statistical Methods, a Primer*, Chapman and Hall, London. 1986.

[19] C. Nastar and N. Ayache. Non-Rigid Motion Analysis in Medical Images: A Physically Based Approach. Procs of the International Conference on Information Processing in Medical Images, pp 17-32, Springer Verlag, 1993.

[20] C.Nastar, B.Moghaddam and A.Pentland. Generalized Image Matching: Statistical Learning of Physically-Based Deformations. Procs ECCV96 (eds. B.Buxton and R.Cipolla), 1, pp 589-598, Springer, 1996.

[21] A. Samal and P. Iyengar. Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey. Pattern Recognition, Vol. 25, no. 1, pp. 65-77, 1992.

[22] P.D. Sozou, T.F. Cootes, C.J. Taylor, E.C. Di Mauro. Non-Linear Point Distribution Modelling Using a Multi-Layer Perceptro, Procs of the British Machine Vision Conference, ed. David Pycock, BMVA Press, vol 1, pp 107-116, 1995.

[23] D. Terzopoulos, K. Waters. Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models. IEEE Trans. of PAMI, vol 15, no 6, pp 569-579, 1993.

[24] M. Turk and A. Pentland. "Eigenfaces for Recognition". Journal of Cognitive Neuroscience, Vol 3, no 1, pp 71-86, 1991.

[25] D. Valentin, H.Abdi, A. O'Toole and G.W. Cotrell. Connectionist Models of Face Processing: A Survey. Pattern Recognition, Vol. 27, no. 9, pp. 1209-1230,1994.

[26]A.L. Yuille, D.S. Cohen and P. Halliman. "Feature Extraction From Faces Using Deformable Templates". International Journal of Computer Vision vol 8, pp 104-109, 1992.

IEEE Xplore® RELEASE 1.8

Welcome
United States Patent and Trademark Office

» ABS

Help FAQ Terms IEEE Peer Review

Quick Links

Welcome to IEEE Xplore®

- ○- Home
- ○- What Can I Access?
- ○- Log-out

Tables of Contents

- ○- Journals & Magazines
- ○- Conference Proceedings
- ○- Standards

Search

- ○- By Author
- ○- Basic
- ○- Advanced

Member Services

- ○- Join IEEE
- ○- Establish IEEE Web Account
- ○- Access the IEEE Member Digital Library

IEEE Enterprise

- ○- Access the IEEE Enterprise File Cabinet

Print Format

Search Results [PDF FULL-TEXT 788 KB] PREV NEXT DOWNLOAD CITATION

Request Permissions RIGHTSLINK

Locating faces using statistical feature detectors

Cootes, T.F. Taylor, C.J.
Dept. of Med. Biophys., Manchester Univ., UK;

This paper appears in: **Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on**

Meeting Date: 10/14/1996 - 10/16/1996
Publication Date: 14-16 Oct. 1996
Location: Killington, VT USA
On page(s): 204 - 209
Reference Cited: 14
Number of Pages: xxvi+386
Inspec Accession Number: 5425655

Abstract:

We describe a method of locating hypotheses for the positions of faces in an im use statistical feature detectors to locate candidates for features, then use a s model of the shape and orientation of the features to test combinations of suc to find the most plausible. The best sets can be used as the initial position of a Shape Model, which can then accurately locate the full face

Index Terms:

face recognition image recognition statistical analysis Active Shape Model face loca features full face statistical feature detectors

Documents that cite this document

Select link to view other documents in the database that cite this one.

Search Results [PDF FULL-TEXT 788 KB] PREV NEXT DOWNLOAD CITATION

Home | Log-out | Journals | Conference Proceedings | Standards | Search by Author | Basic Search | Advanced Search | Join IEEE | Web Account | New this week | OPAC Linking Information | Your Feedback | Technical Support | Email Alerting | No Robots Please | Release Notes | IEEE Online Publications | Help | FAQ| Terms | Back to Top

Locating Faces Using Statistical Feature Detectors

T.F. Cootes and C.J.Taylor
Department of Medical Biophysics, University of Manchester, Oxford Road,
Manchester. M13 9PT, UK
email: bim,ctaylor@sv1.smb.man.ac.uk

Abstract

Wc dcscribc a method of locating hypotheses for the positions of faces in an image. We use statistical feature detectors to locate candidates for features, then use a statistical model of the shapc and orientation of the features to test combinations of such features to find the most plausible. The best sets can be used as the initial position of an Active Shape Model, which can then accurately locate the full face.

1 Introduction

Deformable models have been shown to be an effective approach for locating faces in images [1,2,5]. Usually the fcatures of the face are found by applying some form of local optimisation to a face model, so a 'good enough' starting approximation is required. Such starting points arc cither supplied by user interaction or obtained in some application-specific manner. This paper proposes a framework for generating hypotheses for all plausible instances of faces in a sccnc. The approach is to determine a set of key features, use statistical feature detectors to locate all examples of these in a scene and then to generate a ranked list of all plausible combinations of the features. We systcmatically consider all possible sets of features, ranking or eliminating each by considering the statistics of the relative positions and oricntation of feature points using statistical shape models [6]. Missing features are dealt with, allowing robustness to occlusion. The best feature sets are then used to instantiate a deformable model known as an Active Shape Model [5,6] which can be run to locate the full face structure. This leads to a generally applicable method, which can locate multiple faces in a given scene.

In the following we will describe the approach in more detail. We will consider related approaches and describe the statistical fcaturc detectors we use. We will then cover the statistical shape models and show how they can bc used to determine how plausible a set of features is. Finally we will show the method working on face images and give results of systematic experiments.

2 Background

Several groups have developed systems for locating faces by locating features such as eyes, noses and mouths in an image. For instance, Yuille *et al* [3] used hand-crafted models and Moghaddam and Pentland [4] use statistical models to locate individual features. Our feature detectors are similar to those of the latter group, but we reinforce the detection of faces by using statistical information about the relative positions of the features.

Yow and Cipolla [7] describe a system for locating faces which uses a gaussian derivative filter to locate candidates for the lines of the eyes, nostrils and mouth, then uses a belief network to model the face shape and select a set of features most likely to be those of the face.

Burl *et al* [8] combine feature detectors with statistical shape models to generate a set of plausible configurations for facial features. Their (orientation independent) feature detectors match template responses with the output of multi-scale gaussian derivative filters. They use distributions of shape statistics given by Dryden and Mardia [9] to test configurations of outputs from detectors which locate possible locations of the eyes and nostrils. They build up hypothesis sets by first considering all pairs of features. From each pair they determinc the regions in which they would expect candidates for other features to lie, and consider all sets of points which lie in these regions. They allow for missing features in thcir derivations, to give robustness to feature detector failure.

Our general approach is similar to that of Burl *et al.* However, we use more complex statistical feature de-

0-8186-7713-9/96 $5.00 © 1996 IEEE

tectors in order to minimise the number of false positive responses. Our detectors are orientation dependent, making them more discriminating. Although they must be run at multiple angles, they return both the position and orientation of the found features. We use a simpler method of calculating the shape statistics to test sets of points, and include the feature orientation in our tests. We systematically calculate all plausible sets of features using a depth first tree search, pruning sub-trees as soon as they become implausible. Our aim is to rapidly determine sets good enough to initialise an Active Shape Model to locate the object of interest accurately.

3 Overview Of Approach

The approach we use is as follows. We assume that we have sets of training images, in which points are labelled on the the faces. In advance we determine a sub-set of points which can be used as features to detect the faces, by considering the performance of feature detectors trained at every point. We train feature detectors for the chosen points, and build statistical models of both the shape of the whole model and that of the sub-set of feature points.

Given a new image, we proceed as follows:

- We find all responses for the feature detectors
- We systematically search through all these responses to determine all plausible sets of features. How plausible a set is is determined by the relative positions of the candidate features and their relative orientations
- We fit the full shape model to the best sub-set, and use this as the starting point for an Active Shape Model, which can then accurately locate the full face.

4 Statistical Feature Detectors

In order to locate features in new images we use statistical models of the grey levels in regions around the features, and use a coarse-to-fine search strategy to find all plausible instances in a new image [13].

4.1 Statistical Feature Models

We wish to locate accurately examples of a given feature in a new image. To deal with variations in appearance of features we use statistical models derived from a set of training examples. These are simply rectangular regions of image containing instances of the feature of interest. For each n_x x n_y patch we sample the image at pixel intervals to obtain an $n = n_x n_y$ element vector, $\mathbf{g}$.

A statistical representation of the grey-levels is built from a set of s example patches, $\mathbf{g}_i$ $(i = 1..s)$. A Principle Component Analysis is applied to obtain the mean, $\bar{\mathbf{g}}$, and t principle modes of variation represented by the n x t matrix of eigenvectors, $\mathbf{Q}$. The value of t is chosen so that the model represents a suitable proportion of the variation in the training set (eg 95%) [6].

Our statistical model of the data is

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{Qc} + \mathbf{r}_g \qquad (1)$$

where the elements of $\mathbf{c}$ are zero mean gaussian with variance λ_i, the elements of $\mathbf{r}$ are zero mean gaussian with variance v_j and the columns of $\mathbf{Q}$ are mutually orthogonal. This is the form of a Factor Model [11], with $\mathbf{c}$ as the common factors and $\mathbf{r}_g$ as the errors.

Given a new example, $\mathbf{g}$, we wish to test how well it fits to the model. We define two quality of fit measures as follows,

$$f_1 = M_t + \frac{R^2}{V_r} \qquad (2)$$

$$f_2 = M_t + \sum_{j=1}^{j=n} \frac{r_j^2}{v_j} \qquad (3)$$

where $M_t = \sum_{i=1}^{i=t} \frac{c_i^2}{\lambda_i}$ $\qquad \mathbf{c} = \mathbf{Q}^T(\mathbf{g} - \bar{\mathbf{g}}) \qquad (4)$

$$R^2 = \mathbf{r}_g{}^T\mathbf{r}_g = (\mathbf{g} - \bar{\mathbf{g}})^T(\mathbf{g} - \bar{\mathbf{g}}) - \mathbf{c}^T\mathbf{c} \qquad (5)$$

$$\mathbf{r}_g = \mathbf{g} - (\bar{\mathbf{g}} + \mathbf{Qc})$$

In [13] we show that the distribution of these fit values is a scaled chi-squared distribution of degree k, $p(f) = (n/k)X^2(kf/n, k)$ where k is the number of degrees of freedom of the pixel intensities. k can be estimated from a verification set of examples; $k = 2(n/\sigma)^2$, where σ is the standard deviation of the distribution of fit values across the set.

The time to calculate f_1 is about half that for f_2, but gives slightly less predictable distributions and poorer discrimination between true positive and false positive responses. Knowledge of $p(f)$ allows us to set thresholds which will produce predictable numbers of false negatives (missed true featurcs). More details of the statistical feature models are given in [13].

4.2 Properties of the Statistical Feature Detectors

The quality of fit of the feature models to an image is sensitive to position, orientation and scalc. By systcmatically displacing the models from their true positions on the training set we can quantify this sensitivity. This information allows us to calculate the number of different scales and orientations at which the detector should be run in order to cover a givcn rangc of object orientation and scales.

4.3 Searching for Features

Given a new image we wish to locate all plausible instances of a given feature. We train feature models at several levels of a gaussian pyramid [12] and determine their sensitivity to angle and scale. We then use a coarse–to–fine search strategy as follows;

- Test every pixel of a coarsc resolution image with the matching feature model at a set of angles and scales. Determine the peaks in response which pass the statistical threshold.
- Refine the position and angle estimates of each response on finer resolution images. The accuracy in angle required is determined by the (pre–computed) sensitivity of the model.

Those which pass the threshold at the finest resolution are candidate features.

5 Use of Statistical Shape Models to Test Hypotheses

When presented with a new image we apply the selected feature detectors over target regions to generate a number of candidates for each feature. By choosing one candidate for each feature we can generate a hypothesis for sets of features belonging to the same object. We will use statistical shape models to determine the plausible sets by considcring the relative positions of the features and the orientations at which each was detected.

5.1 Statistical Shape Models

We have previously described methods for building statistical shape models. Given a training set of shapes, each representing n labelled points, we can find the mean configuration (shape) and the way the points tend to vary from the mean [5,6]. The approach is to align each example set into a common reference frame, represent the points as a vector of ordinates in this frame and apply a PCA to the data. We can use the same formulation as for the grey-level models above,

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Pb} + \mathbf{r} \qquad (6)$$

where $\mathbf{x} = (x_1 \ldots x_n\, y_1 \ldots y_n)^T$, $\mathbf{P}$ is a $2n$ x t matrix of eigenvectors and $\mathbf{r}$ is a set of residuals whose variance is determined by miss–one–out experiments. In this case t is the number of *shape* parameters required to explain say 95% of the shape variation in the training sct.

Again, the quality of fit measure for a new shape is given by

$$f_{shape} = \sum_{i=1}^{t} \frac{b_i^2}{\lambda_i} + \sum_{j=1}^{j=2n} \frac{r_j^2}{v_j} \qquad \begin{array}{l} \mathbf{b} = \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}) \\ \mathbf{r} = \mathbf{x} - (\bar{\mathbf{x}} + \mathbf{Pb}) \end{array} \quad (7)$$

Which should be distributed approximately as chi–squared of degree *2n–4*.

In the case of missing points, we can reformulate this test using weights (1.0 for point present, 0.0 for point missing);

$$f_{shape} = \sum_{i=1}^{t} \frac{b_i^2}{\lambda_i} + \sum_{j=1}^{j=2n} w_i \frac{r_j^2}{v_j} \qquad (8)$$

where in this case $\mathbf{b}$ is obtained as the solution to the linear equation

$$(\mathbf{P}^T\mathbf{W})(\mathbf{x} - \bar{\mathbf{x}}) = (\mathbf{P}^T\mathbf{WP})\mathbf{b} \qquad (9)$$

(W is a diagonal weight matrix).

This measure will be distributed as chi–squared of degree $2n_v - 4$ where n_v is the number of points present.

5.2 Models of the Feature Sets

Our features represent a sub–set of the points making up the full shape model for the object of interest. For cach such sub–set we can generate statistical models of the configurations of the feature positions as described above. For instance for the face model we choose features at four of the 122 points of the full model and build statistical models both of the whole set and of the four points. Each shape model has its own co–ordinate frame (usually centred on the centre of gravity of the points and with some suitably normalised scale and orientation [5]).

To test the validity of a set of image points forming a shape, **X**, we must calculate the shape parameters **b** and the pose Q (mapping from model frame to image) which minimise the distance of the transformed model points, **X'**, to the target points

$$\mathbf{X} \approx \mathbf{X}' = Q(\bar{\mathbf{X}} + \mathbf{Pb}) \qquad (10)$$

(Q is a 2D Euclidean transformation with four parameters, t_x, t_y , s and Θ.)

This is a straightforward minimisation problem [6,5]. Having solved for Q and **b** we can project the points into the model frame using Q^{-1} and calculate the residual terms and hence the quality of fit, f_{shape} . We can test the plausibility of the shape probabalistic limits both to the overall quality of fit f_{shape} and, if desired, to the individual shape parameters b_i . The latter have zero mean and a variance of λ_i , the eigenvalues obtained from the PCA.

By considering the training set we can calculate the average mapping between the co-ordinate frame for the full model and that for a sub-set of points. This allows us to propagate any known constraints on the pose of the whole object model to test the pose of the sub-set of points representing the current feature set. In addition, we can learn the expected orientation and scale of each feature relative to the scale and orientation of the set as a whole, allowing further discrimination tests.

If we assume that the configuration of the sets is independent of the errors in feature orientation and scale wc can estimate the probability density for a configuration as follows;

$$p = p(shape) \prod_{i=1}^{i=n_f} [p(\theta_i) p(s_i)] \qquad (11)$$

where the probabilities for shape, angle and scale terms are determined from the estimated distributions. If we assume normal distributions for the measured orientations and scales, then

$$ln(p) = const + \frac{f_{shape}}{2} + \sum_{i=1}^{i=n_f} [\frac{(a_i - \bar{a}_i)}{2\sigma_{ai}^2} + \frac{(s_i - \bar{s}_i)^2}{2\sigma_{si}^2}] \qquad (12)$$

This allows us to sort any plausible hypotheses by their estimated probability.

5.3 Systematic Hypothesis Generation and Testing

If we have n_f feature detectors, and detector i produccs m_i candidates, then there are $\prod m_i$ possible sets. If we allow for the detectors missing true features, then there are $\prod (m_i + 1)$ possible sets (allowing a wildcard feature match). Selecting plausible sets of features given this potential combinatorial explosion has received much attention [8].

We have used a relatively simple scheme amounting to a depth first tree search. The feature candidates are sorted by quality of fit and, if missing features are to be allowed, a wildcard is added. We then recursively construct sets, starting with a candidate from the first detector and adding each candidate from the second detector in turn. The pose and shape of each pair is tested. If a pair passes the tests, each candidate from the third detector is added and the three points tested. Those sets which pass are extended with candidates from the fourth detector and so on. In this manner all possible plausible sets can be generated fairly efficiently. (This approach has the advantage that it can be implemented in a recursive algorithm in a small number of lines of code). We record all the sets which have at least three valid features and pass the statistical tests.

Burl *et al* [8] calculate a probability for each set of candidates which takes into account missing features. We feel that it is difficult to correctly assign probabilities for missing features and instead simply sort our hypotheses first by the number of features present, and secondly by their probability. This avoids comparing sets with different numbers of features directly. In practice it is those which have the fewest missing which tend to be the correct responses.

5.4 Verification of Plausible Feature Sets

Given a plausible sct of fcatures, we find the least-squares fit of the full object shape model to these points. This can be achieved by solving a weighted version of (10), with zero weights for all but the points corresponding to the found features. This gives the starting point for an Active Shape Model. We can run the ASM to convergence, using a multi-resolution search scheme to locate all the points [10]. The ASM has grey-level models of what it expects in the region around every one of its points. By considering the quality of fit ot these models to the image after convergence we can determine whether a good example of the object of interest has been found. Where there are several equally plausible sets of features, the ASM can be run for each and the one with the best final fit accepted. To detect multiple instances of the model the best examples which do not overlap in pose space should be accepted.

6 Results of Experiments

6.1 Locating Facial Features

We have used the system to locate features as a way of initialising a face shape model which can be refined with ASM search. The full model is trained on 122 points marked in each of 40 images (a subset of those used by Lanitis *et al* [14]). Four features based around the eyes and nose were chosen. Figure 1 shows an new face image, the positions of the detected features, thc bcst sct of such features and the position of the full 122 point shape model determined from this set. Figure 2 shows the points after running an ASM to convergence from this starting position. We assumed that we knew the approximate scale, but that the orientation and position were unknown. Each feature detector was run over the whole image and allowed any orientation. It took about 10 seconds on a Sun Sparc20 to run each feature detector over the 512^2 images, then one to two seconds to consider all plausible sets of features. In a real face detection system the orientation is likely to be better constrained, but the features would have to be allowed to vary in scale. Figures 3 and 4 show results for a different image, in which one of the feature detectors has failed to locate a satisfactory candidate. The quality of the full model fit to the found features is worse, but still quite adequate for the ASM to converge to a good solution. This demonstrates the robustness to missing features (and thus to occlusions).

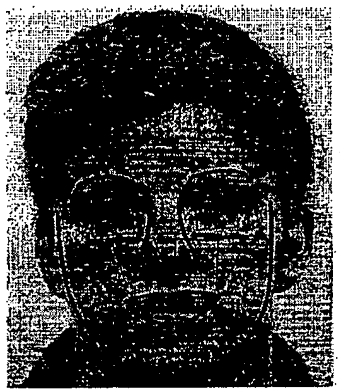


Fig. 1. Candidate fcatures (crosses), best feature set (boxes) and full model fit to the best feature set.

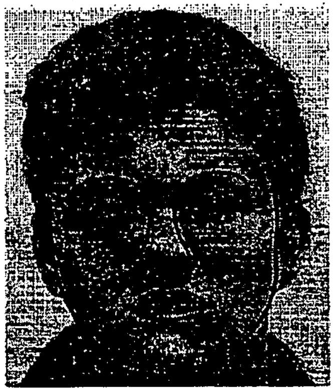


Fig. 2. Full model after running ASM to convergence.

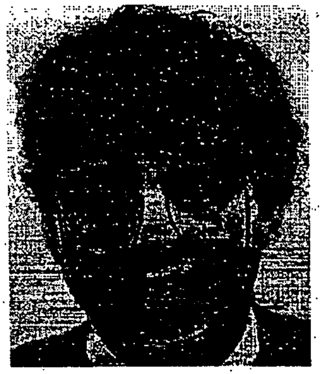


Fig. 3. Candidate features (crosses), best feature set (boxes) (only 3 of 4 found) and full model fit to the best feature set.

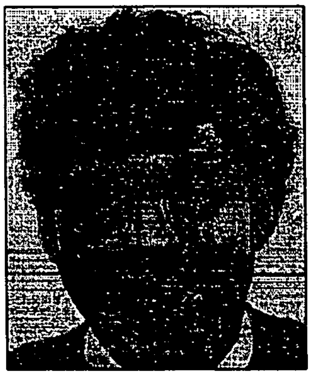


Fig. 4. Full model after running ASM to convergence.

To test the performance more systematically we ran the system on a test set of 40 different images, which had been marked up with the target point positions by hand. On 5 of the images it failed to find any plausible sets, due to multiple feature detector failure. On one image false positives conspired to give a plausible (but wrong) result. On the other 35 the best set gave a good fit. The mean distance between the (known) target points and the full set of points estimated from the best features was 8.7 pixels. The mean error after running the ASM was 6.5 pixels. On average 3.5 of the 4 features were used, and each feature detector found 12 candidates in the image. By more careful choice of size of feature detectors, and by using more detectors (giving more robustness) and a larger training set we expect to improve the results significantly.

7 Discussion

Although we have used one particular form of statistical feature detector, any approach which located the features of interest could be used.

The calculation of plausible feature sets can take a long time if a large number of candidates are to be tested. We are currently interested in generating all plausible hypotheses. However, since the candidates are sorted by quality of fit, we usually find the best overall set quite early in the search. We are currently investigating early verification strategies, terminating the combinatorial search when we find a 'good enough' solution to explain the data.

8 Conclusions

We have shown how statistical feature detectors can be used to find good starting positions for deformable face models, given no prior information on the position or orientation of the object of interest in the image. We used statistical models of the relative positions and orientations of the detected features to determine the plausible sets of features and to limit the possible combinatorial explosion by pruning bad sets as soon as possible. The plausible feature sets can be used to instantiate a statistical shape model representing the whole of the face, which can then be refined using an Active Shape Model. This approach can locate multiple faces in an image.

Acknowledgements

T.F.Cootes is funded by an EPSRC Advanced Fellowship Award.

The face images were gathered and annotated by A. Lanitis.

References

1 M. Kass, A. Witkin and D. Terzopoulos , Snakes: Active Contour Models, in *Proc. First International Conference on Computer Vision*, pp 259–268 IEEE Comp. Society Press, 1987.

2 A. Pentland and S. Sclaroff, Closed-Form Solutions for Physically Based Modelling and Recognition, *IEEE Trans. on Pattern Analysis and Machine Intelligence*. 13, 1991, 715–729.

3. A.L.Yuille, D.S.Cohen and P.Halliman. Feature Extraction From Faces Using Deformable Templates. IJCV vol.8, 1992. pp.104–109.

4. B.Moghaddam, A.Pentland. Probabilistic Visual Learning for Object Detection. *Proc. 5th Int. Conf. on Computer Vision*, 1995. pp. 786–793.

5. T.F.Cootes, C.J.Taylor, D.H.Cooper and J.Graham, Active Shape Models – Their Training and Application. *CVIU* Vol. 61, No. 1, 1995. pp.38–59.

6. T.F.Cootes, A.Hill, C.J.Taylor, J.Haslam, The Use of Active Shape Models for Locating Structures in Medical Images. *Image and Vision Computing* Vol.12, No.6 1994, 355–366.

7. K.C.Yow, R.Cipolla, Towards an Automatic Human Face Localisation System. *in Proc. British Machine Vision Conference*, (Ed. D.Pycock) BMVA Press 1995, pp.701–710.

8. M.C.Burl, T.K.Leung, P.Perona. Face Localization via Shape Statistics. Proc. Int. Workshop on Automatic Face- and Gesture-Recognition (ed. M.Bichsel), 1995. pp. 154–159.

9. I.L.Dryden and K.V.Mardia, General Shape Distributions in a Plane. *Adv. Appl. Prob.* 23, 1991. pp.259–276.

10. T.F.Cootes , C.J.Taylor, A.Lanitis, Active Shape Models : Evaluation of a Multi-Resolution Method for Improving Image Search, *in Proc. British Machine Vision Conference*, (Ed. E.Hancock) BMVA Press 1994, pp.327–338.

11. R.A.Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis. *Prentice–Hall* 1982.

12. P.J.Burt, The Pyramid as a Structure for Efficient Computation. in *Multi–Resolution Image Processing and Analysis*. Ed. Rosenfield, pub. Springer–Verlag. 1984. pp. 6 – 37.

13. T.F.Cootes , G.J.Page, C.B.Jackson, C.J.Taylor, Statistical Grey–Level Models for Object Location and Identification, *in Proc. British Machine Vision Conference*, (Ed. D.Pycock) BMVA Press 1995, pp.533–542. (Also submitted to IVC)

14. A.Lanitis, C.J.Taylor and T.F.Cootes, A Unified Approach to Coding and Interpretting Face Images. Proc. 5th ICCV, IEEE Comp. Soc. Press, 1995, pp. 368–373.